Journal of Official Statistics, vol. 30, n. 1 (2014)

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 1–21, http://dx.doi.org/10.2478/jos-2014-0001

Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models

Jorre T.A. Vannieuwenhuyze[1], *Geert Loosveldt*[2], *and Geert Molenberghs*[3]

The confounding of selection and measurement effects between different modes is a disadvantage of mixed-mode surveys. Solutions to this problem have been suggested in several studies. Most use adjusting covariates to control selection effects. Unfortunately, these covariates must meet strong assumptions, which are generally ignored. This article discusses these assumptions in greater detail and also provides an alternative model for solving the problem. This alternative uses adjusting covariates, explaining measurement effects instead of selection effects. The application of both models is illustrated by using data from a survey on opinions about surveys, which yields mode effects in line with expectations for the latter model, and mode effects contrary to expectations for the former model. However, the validity of these results depends entirely on the (ad hoc) covariates chosen. Research into better covariates might thus be a topic for future studies.

Key words: Selection effects; measurement effects; back-door model; front-door model; causal inference; opinion about surveys.

1. Introduction

Mixed-mode surveys are becoming increasingly popular for the collection of data from general populations (De Leeuw 2005, Voogt and Saris 2005, Dillman et al. 2009b, Vannieuwenhuyze and Loosveldt 2013). A *mixed-mode survey* is a survey in which data from different sample units is collected by different (sets of) data-collection modes. These include Computer-Assisted Personal Interviewing (CAPI), Computer-Assisted Telephone Interviewing (CATI), Postal Self-Administered Questionnaires (Postal SAQs), or Web Self-Administered Questionnaires (Web SAQs). The sample units can be defined either as individual sample members in cross-sectional data, or as time points within individual sample members in longitudinal surveys, so that each sample member is represented by different units.

Sample units can be selected for the data collection modes in four ways. First, in a sequential design, the modes are offered sequentially during a series of contact attempts. Second, in a concurrent design, all the modes are offered simultaneously during the first contact attempt and the sample members choose their preferred mode for responding.

[1] Institute for Social & Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom. Email: jtavan@essex.ac.uk
[2] Centre for Sociological Research, KU Leuven, Parkstraat 45, Leuven 3000, Belgium. Email: geert.loosveldt@soc.kulueven.be
[3] I-BioStat, KU Leuven, Leuven, and Universiteit Hasselt, Diepenbeek, Belgium. Email: geert.molenberghs@med.kulueven.be

Third, in a comparative design, sample units are allocated to data-collection modes on the basis of some stratifying characteristics (for example, different countries use different modes in cross-national surveys, non-Internet households are approached by post instead of a web questionnaire, or different modes are used during different waves in a longitudinal survey). Fourth, in an allocative design, sample units are allocated to the data-collection modes in an experiment-wise random manner (however, each sample member can still choose whether or not to respond to the allocated mode).

Mixed-mode surveys are argued to have advantages over single-mode surveys because they may produce lower selection error, that is, the error introduced by only observing a small subset of the population instead of the entire population (De Leeuw 2005, Voogt and Saris 2005). First, a mixed-mode survey may reduce systematic selection error (e.g., nonresponse error or coverage error) compared to a single-mode survey, because certain members of the population might not be willing or able to respond to the mode used in the single-mode survey, but might respond to an alternative mode in the mixed-mode survey. In this case, the mixed-mode survey offers greater external validity than the single-mode survey. Second, a mixed-mode survey might reduce random selection error (e.g., sampling error) because some respondents may respond through a comparatively low-cost mode in a mixed-mode survey whereas the data-collection cost per unit would be higher in a single-mode survey. As a result, larger samples can be obtained within the same budget constraints. In this case, the mixed-mode survey offers greater external reliability than the single-mode survey.

As a consequence of the lower selection error, mixed-mode surveys provide, on average, samples that represent the population better compared to single-mode surveys, and thus parameter estimates that are closer to the population parameter or have smaller standard errors. However, it should be noted that the argument of lower selection error starts from the assumption that people's willingness to respond in a single-mode survey would persist in a mixed-mode survey that includes the same mode. This assumption might not hold in all situations because, for example, some studies observed lower response rates in a concurrent web and postal mixed-mode design compared to its postal only single-mode counterpart (Medway and Fulton 2012, Millar and Dillman 2011). Nevertheless, this assumption is further considered true throughout this article and we ignore situations where this assumption does not hold true.

Nevertheless, a necessary condition in order for mixed-mode surveys to obtain better representing samples is a *selection effect* between the modes, which means that sample units selected for the different modes differ on the variable of interest (Vannieuwenhuyze et al. 2012). Indeed, if selection effects are absent, then an alternative single-mode design will exist that uses the cheapest mode and provides data of equal external validity but higher external reliability. Evaluating the advantage of mixed-mode surveys thus primarily requires the estimation of selection effects. However, it must be noted that selection effects alone are not sufficient, as will be discussed in Section 5.

Further, evaluating selection effects in mixed-mode data is difficult because they are confounded with another type of *mode effect: measurement effects* (De Leeuw 2005, Voogt and Saris 2005, Dillman et al. 2009b, Weisberg 2005). Measurement effects are differences in measurement error accompanying the different data-collection modes (Voogt and Saris 2005, Weisberg 2005). Measurement effects thus occur when the answers given by the same respondents differ across the modes. As a consequence,

differences between the respondents in the alternate mode groups may either be due to differences in respondent characteristics (a selection effect) or to different measurement of responses (a measurement effect). Measurement effects therefore not only complicate the unbiased estimation of population parameters, but may also counteract the advantages of selection effects with regard to data quality.

The confounding of selection and measurement effects in mixed-mode data overlaps with a central theme of the causal inference literature (see, for example Morgan and Winship 2009, Pearl 2009, Weisberg 2010), which offers two distinct covariate adjustment models for disentangling selection and measurement effects and for obtaining unbiased estimates of population parameters (Pearl 1995, 2009). The first model requires covariates that capture selection effects, while the second model requires covariates that capture measurement effects. To date, both models have scarcely been theoretically discussed in literature relating to mixed-mode surveys. This article aims to fill the gap by providing a thorough theoretical discussion of both models, including the requirements, assumptions, advantages, and disadvantages.

The remainder of the article is structured as follows. Section 2 provides a brief discussion of the causal inference framework, including an overview of formal definitions of the mode effects. Section 3 provides a discussion of both covariate adjustment models and describes the required assumptions and estimation processes. Section 4 provides an illustration of the models using data from a survey about surveys. Section 5 finally concludes the article with a number of important suggestions for future research.

2. The Problem of Counterfactuals

For simplicity, this article is restricted to situations with only two data-collection modes, which we refer to as m_1 and m_2. Further, the article is also restricted to the estimation of the population mean μ on a variable of interest Y. Expansion into situations with more than two modes and more complex parameters can be derived straightforwardly from the following explanation, but may require more complex analysis frameworks.

The occurrence of measurement effects between modes means that the mode has a causal effect on the variable of interest and that respondents would have responded differently if different data-collection modes had been used. As a result, two potential outcomes are theoretically defined for each sample unit in which each potential outcome reflects the unit's outcome on variable Y if one particular mode had been used for data collection (Rubin 1974, Rosenbaum and Rubin 1983). In the general causal inference literature, potential outcomes are traditionally represented on an aggregated level by two different variables, so that each unit is represented by one data line (Holland 1986, Rubin 1974). In this article, by contrast, potential outcomes are represented on a disaggregated level by two different data lines per unit, because such disaggregated representation better allows for uniform definition of mode effects and model assumptions compared to the traditional aggregated representation.

The *full data* thus includes two data lines per sample unit, where each unit's first data line reflects the potential outcome when mode m_1 was used, and the second data line reflects the potential outcome when mode m_2 was used (see Table 1). The full data further requires definition of two additional variables. First, it requires a variable D that indicates

Table 1. The full data includes two data lines per unit, one observed and one counterfactual

Unit U	Mode group G_δ	Mode of data collection D	Potential outcome Y	
1	m_1	m_1	y_{1,m_1}	= observed
1	m_1	m_2	y_{1,m_2}	= counterfactual
2	m_1	m_1	y_{2,m_1}	= observed
2	m_1	m_2	y_{2,m_2}	= counterfactual
3	m_2	m_1	y_{3,m_1}	= counterfactual
3	m_2	m_2	y_{3,m_2}	= observed
4	m_2	m_1	y_{4,m_1}	= counterfactual
4	m_2	m_2	y_{4,m_2}	= observed
⋮	⋮	⋮	⋮	

the distinction between the potential outcomes. This variable is further called the *mode of data collection* and takes the value m_1 or m_2. Second, the full data requires a variable G_δ that indicates the mode for which a unit is actually selected whenever this unit is a sample member of a mixed-mode survey with design δ. This variable is further called the *mode group* and also takes value m_1 when the respondent answers by mode m_1, and m_2 when the respondent answers by mode m_2. It is important to stress that G_δ is design specific. For example, some people may prefer mode m_2 over m_1 in a concurrent design, but would respond by mode m_1 in a sequential design because they are unaware of the subsequent mode m_2.

Nonetheless, within *observed mixed-mode data*, only one data line is observed for each respondent because, by definition, all respondents in mode group m_1 complete the survey by mode of data collection m_1 instead of m_2 and vice versa. Put differently, within mixed-mode surveys, data lines where G_δ and D take different values are not observed (Table 1). For that reason, these data lines are called *counterfactual* (Galles and Pearl 1998, Greenland et al. 1999), but these counterfactuals are, nevertheless, important for the estimation of population means, selection effects, and measurement effects, as will be shown below.

The main objective of a survey is to obtain the best possible estimate of the population mean of the variable of interest. Ideally, the variable of interest is consistently measured over the entire population by one particular mode, which acts as a benchmark. For example, we can use mode m_1 as the benchmark mode, because we believe mode m_1 has a negligible measurement error while mode m_2 is considered to be a distorting mode. As a consequence, the variable of interest is actually defined as $(Y|D = m_1)$ and the population mean is defined as $\mu_{m_1} = E(Y|D = m_1)$, that is, the mean outcome when the values of all population members have been collected by mode m_1. The variable $(Y|D = m_2)$, in contrast, is a biased variable due to measurement error.

Using a mixed-mode design is believed to help obtain a sample that better represents the population, because some population members would not have responded if only one mode had been used, due to particular mode preferences or smaller possible sample sizes. The mixed-mode design thus would provide a better estimate of μ_{m_1}. Nevertheless, unbiased estimation of the population mean μ_{m_1} may still be difficult, because, by the law

of total expectation, it is the weighted sum of two conditional means where one mean requires counterfactual data for estimation:

$$\mu_{m_1} = \mu_{m_1 m_1} \tau_{m_1} + \mu_{m_1 m_2} \tau_{m_2}, \tag{1}$$

where τ_g represents the unconditional probability $P(G_\delta = g)$, and μ_{dg} represents the conditional mean $E(Y|D = d, G_\delta = g)$. The conditional mean $\mu_{m_1 m_1}$ can be estimated from observed mixed-mode data, but the conditional mean $\mu_{m_1 m_2}$ cannot be estimated without additional assumptions because it requires counterfactual data.

Furthermore, the population mean in (1) also clarifies why the estimation of the selection and the measurement effects is of primary interest for the evaluation of mixed-mode data quality. The conditional selection effect on the mean is the difference between the means of the people selected for modes m_1 and m_2 when all responses are measured by the same benchmark mode m_1:

$$S_{m_1}(\mu) = \mu_{m_1 m_1} - \mu_{m_1 m_2}. \tag{2}$$

If this selection effect is zero, then $\mu_{m_1 m_2}$ would be equal to $\mu_{m_1 m_1}$ and to the population mean μ_{m_1}. In this situation, the population mean can be estimated straightforwardly by a single-mode design using mode m_1, which means that a mixed-mode design would be useless for increasing data quality compared to a single-mode design.

The conditional measurement effect on the mean is the difference between the means measured by the two different modes m_1 and m_2 for the same people who are selected for the distorting mode m_2:

$$M_{m_1}(\mu) = \mu_{m_2 m_2} - \mu_{m_1 m_2}. \tag{3}$$

If this measurement effect is zero, then $\mu_{m_1 m_2}$ would be equal to $\mu_{m_2 m_2}$ which can be estimated straightforwardly from the observed mixed-mode data. Put differently, a zero measurement effect would allow unbiased estimation of the population mean μ_{m_1} with mixed-mode data, while a non-zero measurement effect would involve measurement bias on the population mean estimate.

Like the population mean, neither selection nor measurement effects can be estimated without additional assumptions because they require counterfactual data for the estimation of $\mu_{m_1 m_2}$. Indeed, the *overall mode effect*, which is the difference between the directly estimable conditional means of both modes, does not provide any information about the measurement and selection effects as it simply equals their difference, that is,

$$\mu_{m_1 m_1} - \mu_{m_2 m_2} = S_{m_1}(\mu) - M_{m_1}(\mu).$$

Put differently, it is not clear to what extent this difference is caused by a selection effect or a measurement effect. For that reason, selection effects and measurement effects are said to be *confounded* (Morgan and Winship 2009, Pearl 2009).

3. Analysis Models and Assumptions

The previous section made clear that the evaluation of mixed-mode data and the estimation of the population mean require estimation of $\mu_{m_1 m_2}$, which cannot be estimated directly because it requires counterfactual data. The task is to write down this mean in

terms of quantities that can be estimated by observed mixed-mode data, but that require analysis models with assumptions about relations between the variables. This section discusses two possible analysis models which include covariate adjustment.

Before continuing, note that selection and measurement effects are also defined by correlations between variables Y, G_δ, and D (see Figure 1; Pearl 1995, 2009). First, Y may relate to the mode group G_δ due to unobserved common cause variables that simultaneously affect the variable of interest and the mode group for which a respondent is selected (as represented by curved bidirectional edges in Figure 1). The relationship between G_δ and Y thus reflects a selection effect as it implies differences in respondent compositions between the mode groups. Second, by definition, Y is causally affected by the mode of data collection D (as represented by straight unidirectional edges in Figure 1), because the mode defines the measurement error in the response. The effect of D on Y thus denotes the measurement effect between the modes.

In the full dataset, where the responses of all respondents are observed in both modes m_1 and m_2, there is no relationship between D and G_δ (Figure 1a) because two data lines are theoretically defined for each respondent, one for each mode of data collection, irrespective of the actual mode group for which the respondent is selected in the mixed-mode survey. In the observed dataset, in contrast, the mode group G_δ fully determines the mode of administration D for every respondent (as represented by the double-lined edge in Figure 1b) because all respondents in mode group m_1 complete the survey by mode m_1 instead of mode m_2 and vice versa. As a result, G_δ and D are equal and measurement and selection effects are completely confounded.

One could easily proceed by either assuming a zero selection or a zero measurement effect. A zero selection effect would mean that G_δ and Y are unrelated (Figure 1c) and that respondents are completely randomly selected for the different modes. Such random selection overlaps with a proper experimental design and differences between both mode groups would be caused entirely by measurement effects. Nevertheless, a zero selection effect is not only unlikely but also unwanted as discussed in the previous section. A zero measurement effect, in turn, would mean that D and Y are unrelated (Figure 1d), that both modes come with equal measurement error, and that differences between both mode groups are entirely caused by selection effects. Nevertheless, like a zero selection effect, a zero measurement effect is very unlikely within mixed-mode surveys.

Instead of making improbable assumptions about zero selection and measurement effects, the literature about causal inference suggests the inclusion of adjusting covariates into the analysis model (Rosenbaum and Rubin 1983, Rubin 1974). Two types of covariates can be distinguished, where one type controls for selection effects and the other type controls for measurement effects (Pearl 1995, 2009). Both types are discussed in detail throughout the next subsections, which list the required model assumptions and show how both models allow the estimation of the counterfactual mean $\mu_{m_1 m_2}$ if the assumptions hold true.

3.1. *The Back-Door Model*

The first analysis model with covariate adjustment involves the inclusion of a set of covariates B, where B is argued to explain the selection effect as a common cause of Y and

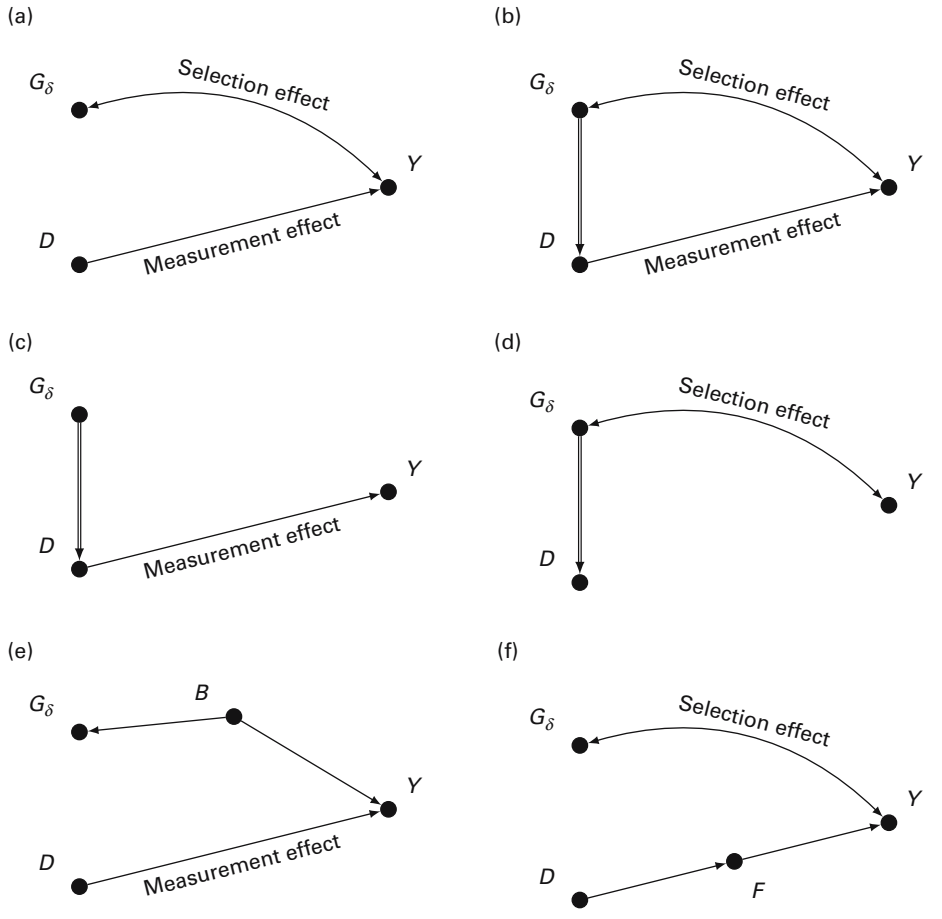

Fig. 1. Relationships between variables in mixed-mode data can be represented by causal graphs, where straight unidirectional edges represent direct causal effects and curved bidirectional edges represent correlations due to unobserved common causes (Pearl 1995, 2009). (a) In the full dataset, the mode group (G_δ) and mode of data collection (D) are independent, and no confounding between measurement and selection effects occurs. (b) In a mixed-mode dataset, the mode group (G_δ) and mode of data collection (D) are equal (double line), and measurement and selection effects are completely confounded. (c) The selection effect is zero when people are completely randomly selected for the different modes. The difference between the mode groups then equals the measurement effect. (d) The measurement effect is zero when all modes introduce equal measurement error. The difference between the mode groups then equals the selection effect. (e) Back-door covariates B allow for unbiased estimation of population means by blocking or explaining the selection effect. (f) Front-door covariates F allow for unbiased estimation of population means by blocking or explaining the measurement effect

G_δ (see Figure 1e). This model is called the backdoor model by Pearl (1995, 2009), because it aims to capture 'back-door' correlations between the survey mode (G_δ) and the variable of interest (Y) which arise from common cause variables.

Nevertheless, the back-door model starts from two assumptions (Pearl 2009, Morgan and Winship 2009). The first assumption is the *ignorable mode selection assumption* and requires that B fully captures the selection effect between the modes or that G_δ and Y are

independent after controlling for B (as represented by the lack of an edge between G_δ and Y in Figure 1e). If this assumption does not hold true, part of the selection effect is not captured and the confounding problem remains. The second assumption is the *mode-insensitivity assumption* and requires the absence of measurement effects on B or that D and B are independent (as represented by the lack of an edge between both variables in Figure 1e). If this assumption does not hold true, part of the measurement effect is channelled through B and the confounding problem again remains. It should, however, be noted that both assumptions cannot be empirically verified, as they refer to differences between observable and counterfactual outcomes.

If both the ignorable mode selection assumption and the mode-insensitivity assumption hold, it can be shown that the counterfactual mean $\mu_{m_1 m_2}$ can be rewritten as an expression of quantities which can be estimated by observed data. For simplicity, let B be a discrete variable, μ_{dgb} represent the conditional mean $E(Y|D = d, G_\delta = g, B = b)$, and $\pi_{b|dg}$ represent the conditional probability $P(B = b|D = d, G_\delta = g)$. The following result emerges:

$$\begin{aligned} \mu_{m_1 m_2} &= \sum_b \mu_{m_1 m_2 b}\, \pi_{b|m_1 m_2} \\ &= \sum_b \mu_{m_1 m_2 b}\, \pi_{b|m_2 m_2}. \end{aligned} \tag{4}$$

The first step of (4) is an application of the law of total expectation. The second step follows from both assumptions. Indeed, $\mu_{m_1 m_2 b} = \mu_{m_1 m_1 b}$ because $Y \perp G_\delta|(D, B)$ by the ignorable mode selection assumption, and $\pi_{b|m_1 m_2} = \pi_{b|m_2 m_2}$ because $B \perp D|G_\delta$ by the mode-insensitivity assumption. As a result, implementing (4) into (1), (2), and (3) allows estimation of the population mean, the selection effect, and the measurement effect once an appropriate set of back-door variables is available:

$$\begin{aligned} \mu_{m_1} &= \sum_b \mu_{m_1 m_1 b}\left(\pi_{b|m_1 m_1}\, \tau_{m_1} + \pi_{b|m_2 m_2}\, \tau_{m_2}\right), \\ S_{m_1}(\mu) &= \sum_b \mu_{m_1 m_1 b}\left(\pi_{b|m_1 m_1} - \pi_{b|m_2 m_2}\right), \\ M_{m_1}(\mu) &= \sum_b \pi_{b|m_2 m_2}\left(\mu_{m_2 m_2 b} - \mu_{m_1 m_1 b}\right). \end{aligned} \tag{5}$$

Within the existing literature concerning causal inference, the back-door model is widely known due to the seminal work of Rubin (2005, 1991, 1978, 1974). Nevertheless, within Rubin's framework, the ignorable mode-selection assumption is formulated thoroughly, but the mode-insensitivity assumption is formulated less than clearly by the mere requirement that covariates must be collected at baseline (that is before treatment in an experimental study). As a result, within the existing literature concerning mixed-mode survey data, the back-door model has already been widely applied (see, for example Lugtig et al. 2011, Heerwegh and Loosveldt 2011, Jäckle et al. 2010, Hayashi 2007, Fricker et al. 2005, Holbrook et al. 2003, Greenfield et al.

2000), but most of these studies use sociodemographic variables as back-door covariates. Such variables might easily be argued to be mode-insensitive, but they might not sufficiently explain why different people are selected for the different modes (Vannieuwenhuyze and Loosveldt 2013). Nonetheless, this issue is largely ignored within existing studies. Future studies might therefore focus on the search for better back-door covariates, such as paradata or survey questions asking for mode preferences (see, for example Olson et al. 2012).

3.2. The Front-Door Model

The second analysis model with covariate adjustment involves the inclusion of a set of variables F, where, in contrast to the back-door model, F is argued to explain the measurement effect as an intermediate variable between Y and D (see Figure 1f). This model is called the front-door model by Pearl (1995, 2009), because it aims to capture 'front-door' correlations between the survey mode and the variable of interest which arise from a direct causal effect of the survey mode (D) on the variable of interest (Y).

Like the back-door model, the front-door model also starts from two assumptions (Pearl 2009, Morgan and Winship 2009). The first is the *exhaustiveness assumption* and requires that F fully captures the measurement effects between the modes or that D and Y are independent after controlling for F (as represented by the lack of an edge between F and Y in Figure 1f). If this assumption does not hold true, part of the measurement effect is not captured and the confounding problem remains. The second assumption is the *isolation assumption* and requires the absence of selection effects on F or that G_δ and F are independent (as represented by the lack of an edge between both variables in Figure 1f). If this assumption does not hold true, part of the selection effect is channelled through F and the confounding problem again remains. However, it should be noted that as with the back-door model, both assumptions cannot be empirically verified as they refer to differences between observable and counterfactual outcomes.

Similarly to the back-door model, if both the exhaustiveness assumption and the isolation assumption hold true, it can be shown that the counterfactual mean $\mu_{m_1m_2}$ can be rewritten as an expression of quantities which can be estimated by observed data. For simplicity, let F be a discrete variable, μ_{dgf} represent the conditional mean $E(Y|D=d, G_\delta=g, F=f)$, and $\pi_{f|dg}$ represent the conditional probability $P(F=f|D=d, G_\delta=g)$. The following result emerges:

$$\begin{aligned}\mu_{m_1m_2} &= \sum_f \mu_{m_1m_2f}\ \pi_{f|m_1m_2} \\ &= \sum_f \mu_{m_2m_2f}\ \pi_{f|m_1m_1}.\end{aligned} \tag{6}$$

Once again, the first step of (6) is an application of the law of total expectation, while the second step follows from both assumptions. Indeed, $\mu_{m_1m_2f} = \mu_{m_2m_2f}$ because $Y \perp D|(G_\delta, F)$ by the exhaustiveness assumption, and $\pi_{f|m_1m_2} = \pi_{f|m_1m_1}$ because $F \perp G_\delta|D$ by the isolation assumption. As a result, implementing (6) into (1), (2), and (3) allows estimation of the population mean, the selection effect, and the measurement effect once

an appropriate set of front-door variables is available:

$$\begin{aligned}\mu_{m_1} &= \sum_f \pi_{f|m_1m_1}\left(\mu_{m_1m_1f}\,\tau_{m_1} + \mu_{m_2m_2f}\,\tau_{m_2}\right),\\ S_{m_1}(\mu) &= \sum_f \pi_{f|m_1m_1}\left(\mu_{m_1m_1f} - \mu_{m_2m_2f}\right),\\ M_{m_1}(\mu) &= \sum_f \mu_{m_2m_2f}\left(\pi_{f|m_2m_2} - \pi_{f|m_1m_1}\right).\end{aligned} \tag{7}$$

Even though the front-door model is analytically a mirror image of the back-door model, it is hardly mentioned in the literature on causal inference and we have found no mention to date in the literature on mixed-mode surveys. The front-door model requires variables that explain why people respond differently in the different modes. Therefore, front-door variables should try to measure, among other items, response burdens, satisficing, acquiescence, or social desirability. Potential front-door variables might be questions about, among others, survey pleasure or survey experiences (see, for example Loosveldt and Storms 2008), or variables including information about the number of item nonresponses or primacy and recency effects. For example, in Section 4, a question is used about whether the respondents found answering the survey a pleasant or unpleasant task. This variable provides results in line with expectations, even though it was selected ad hoc because the data was not collected with the idea of using the front-door model. The front-door model also therefore requires future research on the development and operationalisation of better front-door covariates.

4. An Illustration Using Data from a Survey About Surveys

4.1. Data Collection

The application of the back-door and front-door models will be illustrated by using them in connection with data from a survey concerning opinions about surveys, which was organised in 2004 in Flanders, Belgium, by the Survey Methodology Research Group of the Centre for Sociological Research, KU Leuven (Storms and Loosveldt 2005). The total sample consisted of 960 Flemish people aged between 18 and 80, sampled from the national register. A two-stage sampling procedure was used in which 48 communities were first selected with probability proportional to size and with replacement. Subsequently, 20 people were randomly drawn from each selected community. The clustering within communities is taken into account in the analyses and the data is weighted for differential nonresponse rates within the communities to preserve equal cluster sizes. Within-cluster nonresponse is further assumed to be ignorable.

A sequential mixed-mode design was used to collect the data (Figure 2). Each sample member was first contacted by post with an invitation to complete an enclosed paper questionnaire. If a sample member did not return the postal questionnaire, a first reminder was sent two weeks later and a second reminder accompanied by a new questionnaire was sent four weeks after the first reminder. The postal survey phase lasted two months in total. Sample members who did not return the paper questionnaire in due time were contacted by

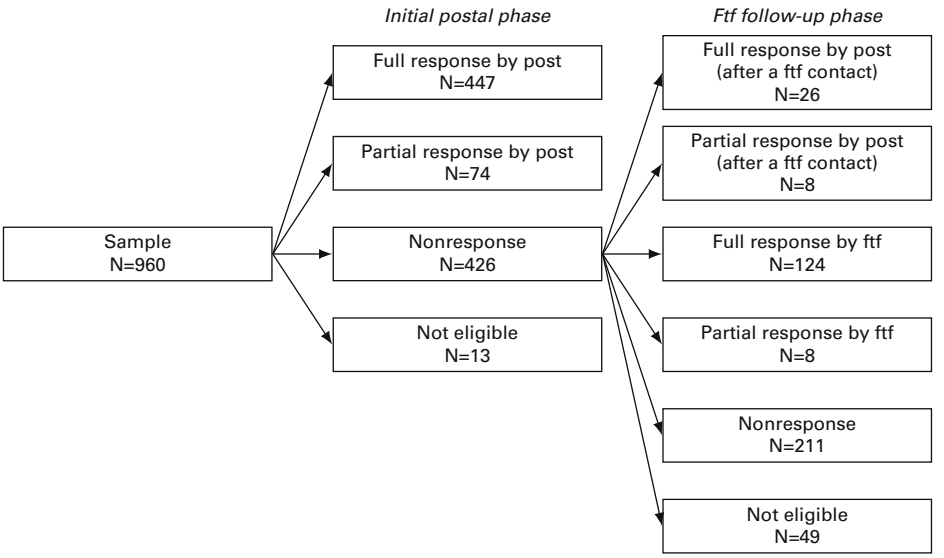

Fig. 2. The survey about surveys used a sequential mixed-mode design starting with a postal phase and ending with a face-to-face (ftf) follow-up

an interviewer at home to complete a face-to-face interview (i.e., CAPI). This face-to-face follow-up was not made known to the sample members during the initial postal phase.

For simplicity, the analyses will only include those respondents who responded to all the variables listed below. Only considering full responses, the initial postal phase reached a response rate (= full response/total sample − not eligible) of 47.20%, which the face-to-face follow-up increased to 63.04% (Figure 2). This response rate is relatively high for a general population survey.

4.2. Variables

4.2.1. Variables of Interest

Mode effects are analysed on the means of six items, each measuring a certain dimension of a short scale representing the respondents' opinions about surveys (Loosveldt and Storms 2008). These items include statements about whether surveys are useful, whether surveys are a waste of people's time, whether surveys stop people doing more important things, whether surveys are boring for respondents, whether the respondent likes surveys, and whether surveys are an invasion of privacy (Table 2). Respondents could indicate agreement or disagreement with these statements on a 5-point Likert scale ranging from 'completely disagree' to 'completely agree'. In the postal questionnaire, these answer categories were listed horizontally in a table, but a 'don't know'/'no opinion' option was not provided. In the face-to-face interviews, the response categories were read out by the interviewer and presented vertically on a showcard, again excluding 'don't know' and 'no opinion' options. For the analyses, all items were rescaled so that high values indicate positive opinions and low values indicate negative opinions.

Table 2. The survey about surveys includes six items/statements about surveys (Loosveldt and Storms 2008). Each respondent could indicate agreement or disagreement with each statement on a 5-point Likert scale (completely disagree, disagree, neither agree nor disagree, agree, completely agree)

Var.	Description
Y_1	'Surveys are useful ways of gathering information.'
Y_2	'Most surveys are a waste of people's time.'
Y_3	'Surveys stop people doing more important things.'
Y_4	'Surveys are boring for the persons who have to answer the question.'
Y_5	'I do not like participating in surveys.'
Y_6	'Surveys are an invasion of privacy.'

The particular topic of the survey might be very likely to cause selection effects and measurement effects on the means. First, there might be selection effects because nonrespondents to the postal questionnaire are likely to be more negative about surveys (Loosveldt and Storms 2008). The postal group data provide some evidence for this expectation: the later a postal questionnaire was returned, the lower the mean opinion score on all six opinion variables (table not included). Second, measurement effects are also expected, because respondents interviewed face-to-face will probably tend to report more positive opinions about surveys (Dillman et al. 2009a, Loosveldt and Storms 2008). Indeed, the mere presence of the interviewer may lead respondents to give socially desirable positive answers that do not reflect the respondents' real opinions.

4.2.2. Back-Door Variables B

The back-door variables include a cross-classification of age and gender, educational level, ownership of a personal email address, activity status, and the number of adults (above 18 years of age), adolescents (between 12 and 18 years of age) and children (under 12 years of age) in the household. Age is divided into six categories, each spanning a period of ten years (18-27, 28-37, 38-47, 48-57, 58-67, and 68-80). The variable for educational level contains six categories: no qualification, primary school, lower secondary, upper secondary, college (non-university), or university. Activity status comprises eight categories: full-time employed, under 50% part-time employed, over 50% part-time employed, unemployed, retired, homemaker, disabled, and 'other'. The numbers of other people in the household also constitute different categories: 1, 2, 3, 4, and 5 or more adults, and 0, 1, and 2 or more adolescents or children.

These variables were chosen because they are very likely to be mode insensitive. Measurement effects are unlikely to occur between a face-to-face interview and a postal questionnaire on variables such as gender, age, the number of household members, or ownership of an email address. Firm evidence for the mode sensitivity of educational level and job-status variables is also lacking within existing literature, even though respondents might tend to overstate their educational attainment and describe themselves as employed when talking to an interviewer because they find these questions embarrassing (Lee and Renzetti 1990, Tourangeau and Yan 2007).

The central question is whether these back-door variables fully capture the selection effect on the variables of interest. Some insights can be provided by regression analysis of the back-door variables on the mode group and the variables of interest. These analyses

indicate significant associations between educational level and the mode group, but no significant associations between the back-door covariates and the variables of interest except for the number of adults and the question about privacy (item Y_6) (tables not included). Although these associations therefore provide little evidence of possible selection effects, they nevertheless neither prove the absence of selection effects nor prove the capturing power of the back-door variables.

For the analyses, the set of back-door variables is transformed into one propensity score variable (Rosenbaum and Rubin 1983, Little 1986, Little and Rubin 2002). The respondents' propensity scores of responding via the postal questionnaire instead of the face-to-face interview are estimated by a maximum likelihood logistic regression model, using the mode group as the dependent variable and the back-door variables as independent variables. Subsequently, the estimated propensities are transformed into a grouped variable by coarsening the propensity scores into five values determined by using the 20th, 40th, 60th, and 80th percentiles as cut points. This coarsened propensity score variable is further used as the back-door variable B.

4.2.3. Front-Door Variable F

As a front-door variable, a question is used which concerns the respondents' experiences during the survey. At the end of the questionnaire, the respondents were asked whether they found answering the questions a pleasant or unpleasant task. The respondents could select an answer from a 5-point Likert scale, comprising 'very pleasant', 'pleasant', 'neither pleasant nor unpleasant', 'unpleasant', and 'very unpleasant'. The format of this question in the postal questionnaire and the face-to-face interview was exactly the same as the opinion about survey items. Because relatively few respondents marked 'very pleasant', 'unpleasant', and 'very unpleasant', the variable was dichotomised ('very pleasant' and 'pleasant' versus 'neither pleasant nor unpleasant', 'unpleasant', and 'very unpleasant').

It is very likely that the mode of data collection has a direct causal effect on responses to the question about survey pleasure. The presence of an interviewer might intensify a feeling of discomfort because the respondent participated although he or she did not fully like the survey. Such a feeling of discomfort is resolved by adapting the reported attitude towards the actual behaviour, that is, by providing a socially desirable answer. Accordingly, the answers on survey pleasure from face-to-face respondents will be positive and consistent with eventual participation. Survey pleasure, in turn, probably has an effect on the reported opinion about surveys because people who report completing the survey as a pleasant task will tend to report more positive opinions about surveys in general.

The central question is whether this front-door variable fully captures the measurement effects on the variables of interest. Some insights can be provided by regression analysis of the mode group on the front-door variable and of the front-door variable on the variables of interest. There is a significant association between the mode group and survey pleasure (table not included). Moreover, even though the face-to-face mode includes more reluctant population members, the face-to-face respondents report a significantly higher pleasure compared to postal respondents. This observation may thus confirm the suggestion of cognitive dissonance. Likewise, the associations between survey pleasure and the opinion

items are always positive and also highly significant. These associations might thus provide some evidence of possible measurement effects. Nevertheless, these analyses neither prove the presence of measurement effects nor prove the capturing power of the front-door variable.

4.3. Estimation Methods

The population means, selection effects, and measurement effects in (5) and (7) are functions of means and proportions which can directly be estimated from the data. The means are estimated by the SURVEYREG procedure in SAS, while the logits of the cumulative versions of the proportions are estimated by the SURVEYLOGISTIC procedure in SAS. These procedures take the clustered nature of the data into account as well as the random sample size of the population subgroups (or domains; see Cochran 1977). These procedures further also provide the covariance matrices of the estimates.

The resulting estimates of the SURVEYREG and SURVEYLOGISTIC procedures are maximum-likelihood estimates, which are known to be asymptotically normal with the mean equal to the population parameter. The Delta method, which uses first-order Taylor-expansions approximations (see Agresti 2002, Casella and Berger 2002, Lehmann 2001), can then be used to derive estimates for the population mean, the selection effects, and the measurement effects. In addition, the Delta method also provides approximate standard errors of the population means, selection effects, and measurement effects estimates, and proves that these estimates are also asymptotically normal.

4.4. Results

The results show remarkable differences between the back-door and the front-door models with respect to the population mean estimates (Table 3). With the back-door model, the means are always larger when measured by a postal questionnaire (μ_{post}) compared to measurement by a face-to-face interview (μ_{ftf}). With the front-door model, the opposite trend is revealed. In contrast to the back-door model estimates, the front-door model estimates are thus in line with the expectation that people represent themselves as more positive about surveys in front of an interviewer due to social desirability bias. Nevertheless, it must be noted that the differences between both modes are mostly small (<0.100 on a 5-point scale).

With respect to selection effects, some differences are also found between the back-door and the front-door models. Taking the face-to-face interview as the benchmark mode (i.e., $S_{\text{ftf}}(\mu)$), the back-door model does not yield large and significant selection effects. The front-door model, in contrast, does yield some significant negative effects. The negative signs of these significant selection effects are also in line with expectations, as they refer to more positive opinions of the postal respondents compared to the face-to-face respondents. The largest selection effect is found on the item about whether the respondent likes surveys (item Y_5). This effect mounts up to -0.57, meaning that, on average, postal respondents rate their liking of survey participation 0.57 higher than face-to-face respondents on a 5-point scale.

Taking the postal questionnaire as the benchmark mode (i.e., $S_{\text{post}}(\mu)$), the back-door model yields one significant negative selection effect for the item about whether surveys

Table 3. The back-door and front-door models provide different estimates with respect to the population mean (μ), selection effects ($S(\mu)$), and measurement effects ($M(\mu)$)

Effect *std.err.*	μ_{ftf}	μ_{post}	$S_{\text{ftf}}(\mu)$	$S_{\text{post}}(\mu)$	$M_{\text{ftf}}(\mu)$	$M_{\text{post}}(\mu)$
Back-door model:						
Y_1	3.650***	3.678***	0.013	−0.065	−0.014	−0.038
	0.088	*0.047*	*0.074*	*0.040*	*0.110*	*0.104*
Y_2	3.066***	3.124***	−0.022	−0.085	−0.027	−0.080
	0.099	*0.059*	*0.092*	*0.049*	*0.127*	*0.122*
Y_3	3.320***	3.360***	−0.021	−0.097*	−0.053	−0.065
	0.094	*0.054*	*0.103*	*0.045*	*0.141*	*0.111*
Y_4	2.991***	3.058***	−0.068	−0.065	−0.038	−0.095
	0.094	*0.052*	*0.073*	*0.043*	*0.113*	*0.112*
Y_5	2.660***	3.015***	−0.155	−0.066	0.180	−0.401**
	0.103	*0.065*	*0.086*	*0.053*	*0.119*	*0.130*
Y_6	3.431***	3.501***	0.089	−0.052	0.100	−0.063
	0.082	*0.048*	*0.088*	*0.039*	*0.126*	*0.095*
Front-door model:						
Y_1	3.678***	3.569***	−0.119	0.073	−0.146***	0.100
	0.070	*0.093*	*0.094*	*0.122*	*0.033*	*0.063*
Y_2	3.098***	3.034***	−0.175	0.028	−0.181***	0.033
	0.089	*0.106*	*0.125*	*0.141*	*0.042*	*0.073*
Y_3	3.348***	3.223***	−0.153	0.076	−0.186***	0.109
	0.075	*0.088*	*0.102*	*0.116*	*0.039*	*0.063*
Y_4	3.023***	2.928***	−0.225*	0.099	−0.195***	0.069
	0.074	*0.094*	*0.108*	*0.125*	*0.044*	*0.063*
Y_5	2.746***	2.654***	−0.569***	0.390**	−0.234***	0.055
	0.095	*0.099*	*0.133*	*0.133*	*0.050*	*0.072*
Y_6	3.482***	3.361***	−0.156	0.125	−0.146***	0.114*
	0.071	*0.078*	*0.097*	*0.103*	*0.034*	*0.058*

****: $p<.001$, **: $p<.01$, *: $p<.05$, the p-values refer to two-sided tests of the null-hypothesis 'parameter=0'. For a description of the variables Y_1 to Y_6, see Table 2.*

stop people doing more important things (Y_3), and the front-door model yields one significant positive selection effect for the item about whether the respondent likes surveys (Y_5). A positive selection effect means that people selected for the postal questionnaire are more positive about surveys than people selected for the face-to-face survey when all data has been measured by the postal questionnaire. The positive front-door estimate for item Y_5 is therefore again in line with expectations, because the face-to-face respondents were nonrespondents to the postal questionnaire. The negative back-door estimate for item Y_3, in contrast, is contrary to expectations.

With respect to measurement effects, the differences between the back-door and front-door models are even more striking. Taking the face-to-face interview as the benchmark mode (i.e., $M_{\text{ftf}}(\mu)$), all back-door estimates are small and insignificant, but the front-door estimates are highly significant and negative. Moreover, all front-door estimates are negative and thus once again in line with expectations. Indeed, negative measurement effects mean that people responding through a postal questionnaire would report more positive opinions when surveyed in a face-to-face interview.

Taking the postal questionnaire as the benchmark mode (i.e., $M_{\text{post}}(\mu)$), the back-door model yields one significant negative selection effect for the item about whether the respondent likes surveys (Y_5), and the front-door model yields one significant positive selection effect for the item about whether surveys are an invasion of privacy (Y_6). Once again, the positive front-door estimate is in line with expectations and the negative back-door estimate is not. Indeed, positive measurement effects here mean that people responding in a face-to-face interview would report less positive opinions when surveyed using a postal questionnaire.

Last, the results also show striking differences between the measurement effects when the postal questionnaire and the face-to-face interview respectively are taken as the benchmark mode ($M_{\text{post}}(\mu)$ and $M_{\text{ftf}}(\mu)$). This difference may point to an interaction effect between measurement error and the mode group. People selected for the postal questionnaire seem to have larger measurement effects between both modes compared to people selected for the face-to-face interview.

4.5. Discussion of the Illustration

To summarise, within the data from the survey examined, there is some evidence of selection effects between the modes, but the relevance of these selection effects may depend on the variable of interest, the analysis model, and on which mode is taken as the benchmark. Significant selection effects may point to a possible advantage of using mixed-mode data collection instead of single-mode data collection. Nevertheless, this advantage might not be guaranteed, because there is also evidence of measurement effects. These measurement effects may counteract the advantage provided by selection effects.

In general, large differences in estimates are observed between the back-door model and the front-door model. It should be emphasised that these differences are not caused by the models themselves, but by the variables that are selected as back-door and front-door covariates. It is very likely that the sociodemographic variables, which are used as back-door covariates, lack sufficient power to explain selection effects on the variables of interest. Further, it also remains unclear how much of the confounding of the selection and measurement effects is reduced by the front-door covariates. Nevertheless, because the front-door results were generally in line with expectations, the front-door covariates seem to perform better than the back-door covariates within this illustration.

5. General Discussion

The main aim of this article was to discuss the use of back-door and front-door models to disentangle selection and measurement effects and to estimate the population mean in mixed-mode survey data. Within relevant existing literature, studies concerning mode effect estimation chiefly use the back-door model, employing sociodemographic variables to explain selection effects. However, such sociodemographic variables probably do not meet the assumptions of the back-door model, which requires that the covariates both are mode insensitive and fully capture the selection effects. The front-door model, by contrast, remains largely unexplored within current literature regarding mixed-mode survey data. This model requires covariates which are assumed to both be insensitive to selection effects and fully capture the measurement effects between the modes.

This article widens the focus beyond the mere theoretical discussion of both the back-door and front-door models and aims to suggest a path for future research. Both the back-door and front-door models are theoretically sound ways of estimating population means, selection effects, and measurement effects, but the practical application of both models might offer challenges because mixed-mode data fit within the framework of so-called enriched data (Molenberghs et al. 2012). Enriched data, like, for example, incomplete data, censored time-to-event data, random-effects models, latent classes, latent variables, or mixture modelling, require strong and often empirically unverifiable assumptions. It is therefore imperative to carefully assemble the broadest possible evidence for the assumptions made in future studies on mixed-mode surveys. These future studies must, however, take the following points into account.

First, actual research on proper back-door as well as front-door covariates is all but nonexistent. Future research must start from other sources. A good source of candidates for back-door covariates might be questions about mode preferences (see, for example, Olson et al. 2012), whilst a good source of candidates for front-door covariates might be questions about survey pleasure or survey experiences (see, for example, Loosveldt and Storms 2008). Another possible source is paradata (see, for example, Kreuter et al. 2010) for both back-door and front-door covariates, but unfortunately the availability of such data might be very mode specific.

Second, the performance of back-door and front-door covariates largely depends on the survey design and the variable of interest. Mode effect estimates depend on the survey design through the mode group variable G_δ, which is design specific. For example, the selection effects and measurement effects in a concurrent mixed-mode design might be different from those in a sequential design. As a consequence, different designs might require different back-door or front-door covariates. Further, mode effect estimates depend on the variable of interest because, for example, lower measurement effects are expected for factual questions than for sensitive questions about opinions. Once again, different kinds of variables of interest might require different back-door or front-door covariates.

Third, there is a need for research on the consequences of departures from the assumptions in both the back-door and the front-door models. Better knowledge of the relationship between the assumptions and mode effects estimation bias might not only help in selecting better covariates, but might also help in selecting optimal survey designs for particular survey topics.

Fourth, even though the back-door and front-door models are presented as two separate models, it should be noted that they can be integrated into the same analysis model. For example, the mode-insensitivity assumption of the back-door model requires the absence of measurement effects on the back-door covariates. Present measurement effects on back-door covariates may, however, be captured by a proper set of front-door covariates. These front-door covariates should not fully explain measurement effects on the variable of interest, but only on the back-door covariates. Likewise, back-door covariates can be used to capture present selection effects on front-door covariates and may guarantee the isolation assumption of the front-door model. The possibility of complex models provides additional opportunities for estimating mode effects and population means. Indeed, some back-door and front-door covariates might not perform well when used separately, but

may do a good job when combined into one analysis model. Nonetheless, it must also be kept in mind that more complex models may lead to estimation and identification problems.

Finally, it should be mentioned that in addition to the back-door and front-door models, a third model exists which also allows for estimation of mode effects. This model makes use of instrumental variables (Bowden and Turkington 1990, Angrist et al. 1996), but requires more complex survey designs and does not allow for estimating all conditional mode effects (Vannieuwenhuyze et al. 2012). Nevertheless, integration of the instrumental variable model, the back-door model, and the front-door model may also provide promising solutions.

Two remarks should be made in conclusion. First, this article describes the analysis of mode effects when only two modes are involved. Nevertheless, both the front-door and back-door models can also be applied when more than two modes are present. In that situation, researchers can use two strategies. In the first, they calculate the selection effects and the measurement effects between the benchmark mode and the other modes separately. In the second, they compare the distorting modes all together at once with the benchmark mode. This latter strategy is justified because the researcher may only be interested in measurement by the benchmark mode, while the separate contribution of the other modes to overall measurement bias is less important.

Second, it was stated in the introduction that the occurrence of selection effects is a primary condition for mixed-mode surveys to be advantageous, but their occurrence is nevertheless not a sufficient condition alone. Indeed, mixed-mode surveys involve higher fixed costs in terms of administration and organisation. An increase in these fixed costs might not be sufficiently compensated for by a decrease in the average cost per sample member through using a mixed-mode design. Especially for small samples, mixed-mode surveys might still not be advantageous over single-mode surveys even though selection effects occur. A cost-benefit analysis comparing mixed-mode and single-mode designs would be appropriate here. Such a cost-benefit analysis, however, first requires the estimation of mode effects and might thus provide a good topic for future studies.

6. References

Agresti, A. (2002). Categorical Data Analysis. Hoboken, NJ: Wiley.

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association, 91, 444–455. DOI: http://www.dx.doi.org/10.1080/01621459.1996.10476902

Bowden, R.J. and Turkington, D.A. (1990). Instrumental Variables. Cambridge: Cambridge University Press.

Casella, G. and Berger, R.L. (2002). Statistical Inference (2nd edition). Duxbury, CA: Pacific Grove.

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

De Leeuw, E.D. (2005). To Mix or not to Mix Data Collection Modes in Surveys. Journal of Official Statistics, 21, 233–255.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B.L. (2009a). Response Rate and Measurement Differences in Mixed-Mode Surveys Using

Mail, Telephone, Interactive Voice Response (IVR) and the Internet. Social Science Research, 38, 1–18. DOI: http://www.dx.doi.org/10.1016/j.ssresearch.2008.03.007

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009b). Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (3rd edition). Hoboken, NJ: Wiley.

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. Public Opinion Quarterly, 69, 370–392. DOI: http://www.dx.doi.org/10.1093/poq/nfi027

Galles, D. and Pearl, J. (1998). An Axiomatic Characterization of Causal Counterfactuals. Foundations of Science, 1, 151–182. DOI: http://www.dx.doi.org/10.1023/A:1009602825894

Greenfield, T.K., Midanik, L.T., and Rogers, J.D. (2000). Effects of Telephone Versus Face-to-Face Interview Modes on Reports of Alcohol Consumption. Addiction, 95, 277–284. DOI: http://www.dx.doi.org/10.1046/j.1360-0443.2000.95227714.x

Greenland, S., Pearl, J., and Robins, J.M. (1999). Causal Diagrams for Epidemiologic Research. Epidemiology, 10, 37–48.

Hayashi, T. (2007). The Possibility of Mixed-Mode Surveys in Sociological Studies. International Journal of Japanese sociology, 16, 51–63. DOI: http://www.dx.doi.org/10.1111/j.1475-6781.2007.00099.x

Heerwegh, D. and Loosveldt, G. (2011). Assessing Mode Effects in a National Crime Victimization Survey Using Structural Equation Models: Social Desirability Bias and Acquiescence. Journal of Official Statistics, 27, 49–63.

Holbrook, A.L., Green, M.C., and Krosnick, J.A. (2003). Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. Public Opinion Quarterly, 67, 79–125. DOI: http://www.dx.doi.org/10.1086/346010

Holland, P.W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81, 945–960. DOI: http://www.dx.doi.org/10.1080/01621459.1986.10478354

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. International Statistical Review, 78, 3–20. DOI: http://www.dx.doi.org/10.1111/j.1751-5823.2010.00102.x

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys. Journal of the Royal Statistical Society, Series A, 173, 389–407. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2009.00621.x

Lee, R.M. and Renzetti, C.M. (1990). The Problems of Researching Sensitive Topics: An Overview and Introduction. American Behavioral Scientist, 33, 510–528.

Lehmann, E.L. (2001). Elements of Large-Sample Theory. New York: Springer.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd edition). London: Wiley.

Loosveldt, G. and Storms, V. (2008). Measuring Public Opinions About Surveys. International Journal of Public Opinion Research, 20, 74–89. DOI: http://www.dx.doi.org/10.1093/ijpor/edn006

Lugtig, P., Lensvelt-Mulders, G.J.L.M., Frerichs, R., and Greven, A. (2011). Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey. International Journal of Market Research, 53, 669–686.

Medway, R.L. and Fulton, J. (2012). When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. Public Opinion Quarterly, 76, 733–746. DOI: http://www.dx.doi.org/10.1093/poq/nfs047

Millar, M.M. and Dillman, D.A. (2011). Improving Response to Web and Mixed-Mode Surveys. Public Opinion Quarterly, 75, 249–269. DOI: http://www.dx.doi.org/10.1093/poq/nfr003

Molenberghs, G., Njeru Njagi, E., Kenward, M.G., and Verbeke, G. (2012). Enriched-Data Problems and Essential Non-Identifiability. International Journal of Statistics in Medical Research, 1, 16–44.

Morgan, S.L. and Winship, C. (2009). Counterfactuals and Causal Inference: Methods and Principles for Social Research. Analytical Methods for Social Research. New York: Cambridge University Press.

Olson, K., Smyth, J.D., and Wood, H.M. (2012). Does Giving People their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination. Public Opinion Quarterly, 76, 611–635. DOI: http://www.dx.doi.org/10.1093/poq/nfs024

Pearl, J. (1995). Causal Diagrams for Empirical Research. Biometrika, 82, 669–688. DOI: http://www.dx.doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). Causality: Models, Reasoning and Inference (2nd edition). New York: Cambridge University Press.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–55. DOI: http://www.dx.doi.org/10.1093/biomet/70.1.41

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology, 66, 688–701. DOI: http://www.dx.doi.org/10.1037/h0037350

Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics, 6, 34–58.

Rubin, D.B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. Biometrics, 47, 1213–1234.

Rubin, D.B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association, 100, 322–331. DOI: http://www.dx.doi.org/10.1198/016214504000001880

Storms, V. and Loosveldt, G. (2005). Procesevaluatie van het Veldwerk van een Mixed Mode Survey naar het Surveyklimaat in Vlaanderen. Leuven: KUL, Centrum voor Sociologisch Onderzoek.

Tourangeau, R. and Yan, T. (2007). Sensitive Questions in Surveys. Psychological Bulletin, 133, 859–883.

Vannieuwenhuyze, J.T.A. and Loosveldt, G. (2013). Evaluating Relative Mode-Effects in Mixed Mode Surveys: Three Methods to Disentangle Selection and Measurement

Effects. Sociological Methods and Research, 42, 82–104. DOI: http://www.dx.doi.org/10.1177/0049124112464868

Vannieuwenhuyze, J.T.A., Loosveldt, G., and Molenberghs, G. (2012). A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. International Statistical Review, 80, 306–322. DOI: http://www.dx.doi.org/10.1111/j.1751-5823.2011.00167.x

Voogt, R.J. and Saris, W.E. (2005). Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. Journal of Official Statistics, 21, 367–387.

Weisberg, H.F. (2005). The Total Survey Error Approach: A Guide to the New Science of Survey Research. Chicago: University of Chicago.

Weisberg, H.F. (2010). Bias and Causation: Models and Judgment for Valid Comparisons. Hoboken, NJ: Wiley.

Received April 2012
Revised April 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 23–43, http://dx.doi.org/10.2478/jos-2014-0002

Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey

Kea Tijdens[1]

Occupation is key in socioeconomic research. As in other survey modes, most web surveys use an open-ended question for occupation, though the absence of interviewers elicits unidentifiable or aggregated responses. Unlike other modes, web surveys can use a search tree with an occupation database. They are hardly ever used, but this may change due to technical advancements. This article evaluates a three-step search tree with 1,700 occupational titles, used in the 2010 multilingual *WageIndicator* web survey for UK, Belgium and Netherlands (22,990 observations). Dropout rates are high; in Step 1 due to unemployed respondents judging the question not to be adequate, and in Step 3 due to search tree item length. Median response times are substantial due to search tree item length, dropout in the next step and invalid occupations ticked. Overall the validity of the occupation data is rather good, 1.7-7.5% of the respondents completing the search tree have ticked an invalid occupation.

Key words: Job title; CAWI; occupation database; ISCO; paradata; time stamps; respondent's interest; respondent's age and education; total survey dropout; validity.

1. Introduction

The increasing popularity of web surveys as a new mode of data collection has fundamentally challenged traditional survey methodology. This article focuses on one feature of web surveys, namely how web surveys can substitute the absent interviewer for the survey question concerning occupations. Occupation is a key variable in socioeconomic research, used in studies on labour force composition, social stratification, gender segregation, skill mismatch, and many others. In web surveys the question about occupation is judged risky, as is for example noted by Statistics Netherlands in an exploration of the use of web surveys for their Labour Force Survey (Van der Laan and Van Nunspeet 2009). The authors' worries relate to, among others, breaks in the time series in the measurement of occupations due to the use of different survey modes. They aim to make improvements before using a web survey for their Labour Force Survey. In September 2011, Eurostat organised a workshop on data collection for social surveys using multiple modes, focusing on the measurement of occupations in web surveys among

[1] University of Amsterdam, Amsterdam Institute for Advanced Labour Studies (AIAS), Postbus 94025 1090 GA, Amsterdam, The Netherlands. Email: K.G.Tijdens@uva.nl

Acknowledgments: This article builds on research work done for the EU-funded FP6 project EurOccupations (no 028987, 2006-2009, www.euroccupations.org) and for the *WageIndicator* web survey on work and wages (www.wageindicator.org). The author thanks the editor and three anonymous referees for their valuable comments on earlier versions of this article. The WageIndicator Foundation is gratefully acknowledged for providing access to their database and Maarten van Klaveren for language editing. The author would like to acknowledge the contribution of WEBDATANET [COST Action IS1004].

others. As in other survey modes, most web surveys use an Open-Ended Question (OEQ) for the occupation question. Yet several drawbacks are associated with this OEQ, as will be discussed in this article.

A unique feature of web surveys is that they allow for a closed survey question on occupation, using a search tree and an underlying database of occupations. Despite this, search trees are hardly used in web surveys, although recent techniques such as text string matching, single page filtering and Application Programming Interface (API) as well as an increasing use of multi-country surveys may favour the use of a closed survey question in web surveys over that of an OEQ. This stresses the need for a data quality assessment of occupation search trees in web surveys that is not available to date. This article investigates the dropout rate, the response time and the validity of the ticked occupation in a search tree in the continuous, worldwide *WageIndicator* web survey, using the World database of occupations (WISCO) designed by the author for use in this web survey. Section 2 reviews the ISCO international occupational classifications and the pros and cons of the measurement of occupations in web surveys (CAWI) and in the three other survey modes (PAPI, CATI, and CAPI). This section also details the WISCO search tree and database of occupations. Section 3 reviews explanations for dropout rates and response times, presents hypotheses and details the data used. The results of the analyses concerning the dropout rates during search tree completion, the response time and the validity of the occupation data are discussed in Section 4. The article ends with conclusions and discussion (Section 5).

2. Reviewing the Measurement of Occupations

2.1. The ISCO Occupational Classification

A number of industrialised countries have their own occupational classifications, such as the US, the UK, Germany, the Netherlands, and France. To facilitate cross-country comparisons, Eurostat requires the National Statistical Offices of the EU countries to deliver the occupation variable in their labour force data using the International Standard Classification of Occupations (ISCO). For more than half a century the ISCO has been issued by the International Labour Organisation (ILO), a United Nations organisation (Hunter 2009). ISCO provides a hierarchical classification system with four levels. The ISCO-08 update is increasingly being adopted worldwide. The European Union (2009) has adopted ISCO-08 as its occupational classification.

In ISCO-08 job titles with the same set of tasks and duties performed by one person are aggregated into 433 ISCO four-digit occupation units, which on the basis of similarities of tasks and duties are grouped into three- and two-digit groups. In turn, the latter are grouped into nine one-digit groups on the basis of four skill levels (Greenwood 2004). Although Eurostat has gone to great effort to encourage cross-country discussions about coding problems, an empirical underpinning of the similarity of occupation coding across countries is still lacking. The more disaggregated the hierarchical level, the larger the problem. Elias and McKnight (2001) identify several problems in multi-country datasets and call for the harmonisation of survey questions, the adoption of common coding procedures and a common understanding of the conceptual basis of ISCO, in particular its

skill concept. They stress the need to undertake studies for validity testing of occupations measurement. Apart from the Eurostat discussion platform for National Statistical Offices, hardly any cross-country studies have investigated whether similar job titles are coded into the same ISCO-08 four-digit level.

2.2. The Open Response Formats in PAPI, CATI, CAPI and CAWI

Many socioeconomic surveys, such as Labour Force Surveys (LFS) and Censuses, include a question "What is your occupation?", "What kind of work do you do?" or similar, using either an open or a closed response format. Both formats can be used in all four survey modes, but the Open-Ended Question (OEQ) is most often used. Ganzeboom (2010, p. 7) advises using the open format, "because occupations are complicated". Compared with the variables education and industry, which are also mostly asked in an open response format, the measurement of occupations is problematic given that in many countries the stock of job titles may exceed 100,000 and that the occupational distribution has a very long tail, challenging the number of categories in a coding index or lookup database. For example, the state of Texas, USA, reported over 500,000 job titles in its job evaluation system (Tippins and Hilton 2010). In the OEQ respondents report their job titles as they like, implying that the data collector has to code the job titles according to a national or international occupational classification. CAPI and CATI allow for field and office coding, but PAPI and CAWI have to rely solely on office coding. (Semi-)automatic indexes can be used to assign occupational codes.

The response to the occupation OEQ varies largely. Respondents tend to report their job title in great detail, as they know it from their employment contract, a job evaluation scheme, or a common understanding in the workplace, but they may also report highly aggregated categories, such as 'clerical worker' or 'teacher', or unspecific categories, such as 'employee of department X' or 'senior supervisor'. In CAPI or CATI interviewers will prevent ambiguous, crude or overly detailed responses, but in PAPI and CAWI this is not the case. In CAWI the share of inadequate answers may be even larger than in PAPI, taking into account the habit of web visitors to key in whatever they like. Ganzeboom (2010) suggests coding crude titles in ISCO one- or two-digits, using trailing zeroes. He concludes that office coding can lead to substantial percentages of unidentifiable responses and to data at various levels of aggregation. This is confirmed in the World Values Survey, a predominantly postal survey using office coding for the occupation variable. Its 1999 data for Belgium, the Netherlands, and the UK, selecting only respondents with employment status employee or self-employed, reveals that for Belgium the occupation variable is coded only at ISCO88 two-digit and for the Netherlands and the UK also at three- and four-digit (two-digit: NLD 5%, GBR 8%; three-digit: NLD 22%, GBR 26%; four-digit: NLD 72%, GBR 59%; missing BEL 2%, NLD 1%, GBR 6%). Hence the measurement of various levels of aggregation is a much larger problem than the missing values. Note that the data of the unemployed, who are more likely not to be able to report an occupational title, have been excluded in these percentages. In the World Values Survey in Belgium the occupations of the unemployed are coded at two-digit, whereas for the Netherlands and the UK the question is not considered applicable for this group.

Against the backdrop of the wide variety of job titles as well as the occupational dynamics and the organisation specificity of job titles, it is not surprising that Elias (1997) concludes that office coding of occupations is an inexact process. Similarly, Eurostat (2009) states that inconsistencies are large for variables that require codification, such as occupations. In an analysis of the misclassification of occupation descriptions in the US Current Population Survey, Conrad and Couper (2008) find that the longer the occupation description, the less reliably it is coded. Thus a number of arguments call for an exploration of alternatives to the OEQ format.

2.3. *The Closed Response Formats in PAPI, CATI, CAPI and CAWI*

In a closed response format question, a tick list offers respondents a choice of occupational titles for self-identification. This method can be used in all four survey modes. However, in CATI the choice is limited to 5-7 categories that are inevitably highly aggregated. Otherwise the respondents will not remember all items. PAPI allows for a choice of at most 50 categories, because otherwise the printed questionnaire would exceed a reasonable length. CAPI allows for slightly more categories when using show cards. A limited set of choices may result in lower data quality, because it is difficult to assure consistency in how respondents fit their own job titles into the highly aggregated categories, introducing aggregation bias (De Vries and Ganzeboom 2008). This calls for a decomposition of the task, as has been proven to lead to better judgements (Armstrong et al. 1975).

CAWI allows for an almost unlimited choice of occupational titles. To navigate through a large look-up database, a search tree with two or three steps is needed. This so-called multipage filtering is a convenient way to collect data if a variable has too many possible values to be presented on a single page (Funke and Reips 2007). For quite some years now, job sites have used search trees to help web visitors to identify an occupation. In CAWI, an extended search tree is advantageous because aggregation bias and aggregation heterogeneity are prevented and unidentifiable; ambiguous or crude occupational titles are absent. In addition, search trees can easily be applied in multi-country and multi-language surveys, allowing for cross-country comparisons of highly disaggregated occupational data while ensuring comparable survey operations. However, a disadvantage of search trees is that they are cognitively demanding and time-consuming, as will be discussed later.

2.4. *The Web Survey's Occupation Question*

This article analyses the occupation data from the volunteer *WageIndicator* web survey on work and wages, designed by the author (Tijdens et al. 2010). The survey is posted on the national *WageIndicator* websites (www.wageindicator.org). These websites consist of job-related content, labour law and minimum wage information, and a free Salary Check presenting average wages for occupations based on the web survey data. The websites receive millions of visitors because of their collaboration with media groups with a strong internet presence. The first website and its web survey started in the Netherlands in 2001, expanded to other EU member states from 2004 onwards, included countries outside the EU and in other continents from 2006 onwards, and is operational today in 70 countries

in five continents. In return for the free information provided, web visitors are invited to complete the web survey with a lottery prize incentive. The web survey takes approximately ten minutes to complete. Each web survey is in the national language(s) and adapted to the peculiarities of the country. In 2010, 417,137 web visitors started and 134,960 completed the survey, hence a dropout of 68%.

In 2010, the web survey has 22 pages. Page 1 of the *WageIndicator* web survey asks a question about employment status. The main options are employee, self-employed, and unemployed. Pages 2a and 2b ask a few questions of respondents with and without a job respectively. Page 3 asks about region. On pages 4-6, respondents self-identify their occupation by means of a three-step search tree allowing them to navigate through the WISCO multilingual database of occupations with more than 1,700 occupational titles (one page per step). The database details occupations with a greater precision than ISCO-08 four-digit by adding further digits. The closed response format is preferred over an OEQ with office coding. Apart from preventing aggregation bias and aggregation heterogeneity, an OEQ would have required a continuous and costly coding effort for the 70 countries given the large numbers of observations. The long-term experience with the web survey has revealed that respondents like to specify their occupational title, for example supervisor, senior, junior, trainee and similar. To satisfy these respondents, page 7 has a radio button question with these extensions and an OEQ where respondents are invited to add additional text about the occupational title ticked in the search tree. This text data is analysed in Subsection 4.2.

The WISCO database aims to facilitate respondents' easy but valid self-identification of their job title. To do so, the 433 units in the four-digit ISCO-08 classification are certainly too aggregated. A disaggregated list has to optimise between the demand to include as many occupational titles as possible to facilitate valid self-identification and the demand to be as brief as possible to reduce reading time. In WISCO, the aggregation level of occupations is defined as follows: "An occupation is a bundle of job titles, clustered in such a way that survey respondents in a valid way will recognize it as at their job title; an occupation identifies a set of tasks distinct from another occupation; an occupation should have at least a not-negligible number of jobholders and it should not have an extremely large share in the labour force" (Tijdens 2010, p. 16). Following this definition, broad occupational titles with large numbers of jobholders, such as clerk, teacher or nurse, are broken down into disaggregated occupational titles. Where needed, some occupational titles include a reference to industry or firm size, because the occupational coding does not use auxiliary variables. Similarly, handicraft workers have been distinguished from comparable manufacturing workers. For unskilled occupations, broad occupational titles have been preferred, because job holders may perform several jobs in a short period. From the next sections it can be concluded that the database of 1,700 unique occupational titles is sufficiently detailed for the vast majority of respondents in a multi-country survey.

To navigate through the database, WISCO has a three-step search tree based on a clustering of related occupations. The search tree's first step consists of 23 items, using a mixture of broad occupational groups and industry groups, such as 'Agriculture, nature, animals, environment' or 'Care, children, welfare, social work'. The second step specifies the ticked item in the first step and the third step presents the list of occupations related to the choice in the second step. Approximately one fourth of the occupations can be found

through multiple search paths. Screenshots can be seen in Figure. 1, showing that in each step the list of occupations is sorted alphabetically. Due to technical constraints, the web survey uses a one page per step approach with back-and-forth buttons.

All occupational titles in the WISCO database are coded according to the four-digit ISCO-08 classification with follow-up numbers. In reverse, all ISCO-08 four-digit occupational units have at least one entry in the WISCO list of occupations. ISCO-08 has 27 residual ('not elsewhere classified') units, which are useful for office coding but problematic in the case of self-identification. This problem has been solved by rephrasing all 27 residual occupation units as 'Occupational unit X, all other' and sorting them at the bottom of the appropriate third step of the search tree, assuming that respondents have read all occupational titles in that particular step before deciding to tick the residual occupation.

For the multi-country WISCO database, translations by national labour market experts have been preferred over translations by professional translators. The wording of the occupational titles is kept brief, easy to understand, and hopefully unambiguous. Thus the singular is preferred over the plural and beekeeper over apiarist. No different male and female occupational titles have been used, apart from some countries where this was considered necessary. Synonymous titles are not included as these might confuse respondents. If national experts indicated that two distinct occupational titles were not considered distinct in their country, one occupation was removed from the country list. During the preparation of ISCO-08, the main discussions concerned the skill levels assumed with the ISCO one-digit codes (Elias and Birch 2006). In the WISCO country lists of occupations, this skill ambiguity is solved by adding skill requirements to the occupational titles, when known and applicable. For example in Germany, the 'Archivar/in, Diplom (FH)' has been distinguished from the 'Archivar/in, Diplom (Uni)' and the 'Archivar/in, Fachschule'. Skill requirements have been added when national experts

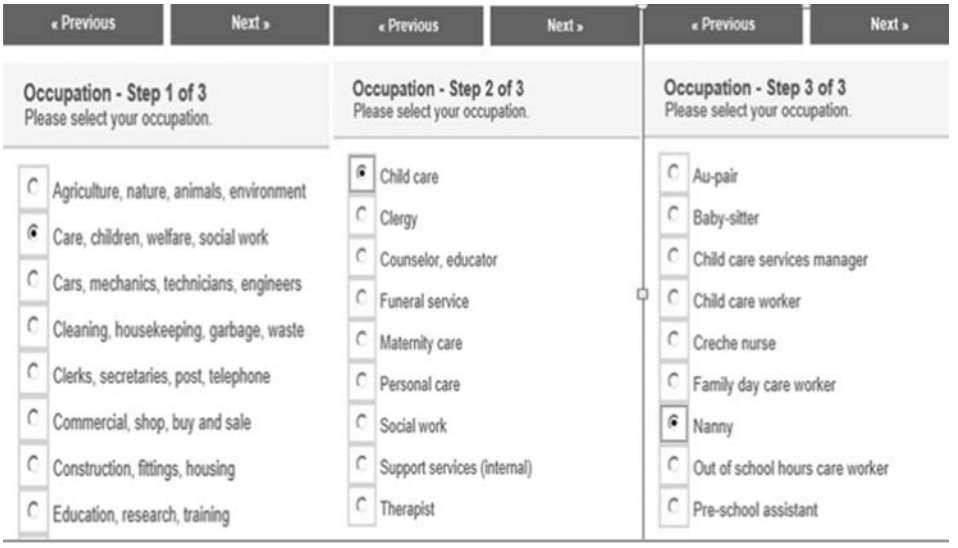

Fig. 1. Screenshots of the pages 4-6 in the occupation survey question in the web survey; note that this figure does not show the full list of 23 entries in Step 1 of 3. Source: *WageIndicator* Survey, UK

indicated a need for it. This turned out to be only relevant in countries where the educational system and the job market are firmly intertwined.

3. Dropout Rates and Response Time

3.1. Explanations for Dropout and Response Time

Given all the efforts to design a database of occupations and a search tree for respondents' self-identification in a web survey, it is certainly important to ask what the response times and dropout rates are, and which theories can explain these outcomes. In addition, how well does the search tree allow respondents to identify their job title as an occupational title from the list, according to the comments posted in an OEQ following the search tree? This section reviews the theoretical explanations and the related hypotheses.

High dropout rates are a major shortcoming of web surveys, threatening data quality. Many studies on dropout rates have been related to the use of progress indicators (e.g., Kaczmirek 2009; Callegaro et al. 2011), but some studies have detailed the impact of respondents' characteristics and survey characteristics on dropout. Dropout is a problem when it is systematic. This might be the case when survey questions are suboptimally formulated, the questionnaire is too lengthy, or other item and survey characteristics are poor (Reips 2002). The support for the interest hypothesis is in line with the findings of Heerwegh and Loosveldt (2006) that personalisation has a significant effect on the probability of starting the web survey and on the probability of reaching and submitting the final web survey page. Galesic (2006) finds in addition that the lower respondents experienced the overall survey burden, the lower the dropout risk. Pages that required more time to complete were followed by dropout more often. Using the German Longitudinal Election Study, Blumenstiel et al. (2010) find that dropout is a function of both respondents' characteristics and page characteristics. Dropout rates are higher for respondents with a lower level of education and in the case of open ended questions. In summary, the length of the questionnaire items, the respondents' level of education and interest in the topic of the questionnaire influence dropout.

Response time has been the subject of increased attention in the survey methodology literature over the last decade. Following the model for analysing survey response proposed by Tourangeau et al. (2000), Yan and Tourangeau (2008) explain response time in web surveys through question complexity and respondents' working memory capacity. They apply a cross-classification model for data from four web surveys in the USA with 27-61 questions. Concerning question complexity, the findings indicate that response times are longer when there are more clauses in a question, more words per clause, larger numbers of answer categories, and more factual and attitudinal questions compared to demographic questions. For respondents' working memory capacity, the authors conclude that the response time is longer for less educated respondents, for older respondents, and for respondents without previous web and survey experience. This is in line with Malhotra's (2008) finding that older respondents take significantly more time to complete a questionnaire. A recent body of knowledge focuses on the impact of question clarity on data quality, including response time. For example, investigating how easily and consistently respondents understand text features in survey questions, Lenzer et al. (2010)

show that the overall effect of seven text features on total response times is highly significant.

This leads us to explore three hypotheses:

Hypothesis 1: We expect that the dropout rates in the occupation search tree are affected by the length of the questionnaire items, operationalised as the number of characters to be read in previous steps of the search tree, and by the respondent's interest in the occupation question, operationalised as the relevance of the question for employed, self-employed and unemployed respondents.

Hypothesis 2: We expect that the response time in each step of the search tree is affected by the search tree item length, the respondent's valid self-identification, the respondent's dropout in the next step, and the respondent's interest, age and education.

Hypothesis 3: We expect that the total survey dropout after the search tree is affected by the response time in the search tree, the respondent's valid self-identification, and the respondent's interest, age and education.

3.2. Data

For the analyses, a new dataset has been compiled, derived from the 2010 second quarter *WageIndicator* web survey in the United Kingdom, Belgium (Dutch), Belgium (French) and the Netherlands. These were the most recent data available at the time of the study. The choice of the three countries was related to the author's language capacities, needed to investigate the respondent's valid self-identification. The new dataset is compiled as follows.

- The web survey contributes data about the ticked items in the 1st, 2nd and 3rd step of the occupation search tree, and the variables employment status, educational level, age, and search tree and total survey dropout.
- The web survey contributes the text that respondents have keyed into the open question following the search tree. The author has coded these responses and the results are shown below. From this data a variable called 'wrong match' is derived, indicating that respondents have keyed in an occupation in the open question other than that ticked in the search tree to identify the validity of the respondent's self-identification.
- The paradata contributes the time stamp for the start of the survey and three time stamps for completion of the 1st, 2nd and 3rd step of the search tree; note that the paradata measures the server-side time stamps, that in case of back-and-forth clicking only the latest time stamps are recorded and that the number of back-and-forth clicks is not recorded.
- The WISCO occupation database contributes the number of characters including blanks and commas in the most efficient search paths for each ticked item in the 2nd and 3rd step, assuming no further reading once respondents have identified their occupations; note that the number of characters read due to back-and-forth clicking is not included.

The total number of observations is 24,811 respondents at the start of the survey, of which 22,990 have completed the survey questions before the search tree and 18,824 have completed the search tree (Table 1). The large majority of respondents are based in

Table 1. Means of respondents' employment status and their education and age in four country/language combinations

	UK	Belgium (French)	Belgium (Dutch)	Netherlands	N
Employee Status					
Employee	0.90	0.88	0.89	0.83	22,990
Self-employed	0.06	0.05	0.04	0.06	22,990
Unemployed	0.04	0.07	0.07	0.11	22,990
Age	35.6	33.2	34.4	35.9	13,194
Education level					
Low education	0.14	0.17	0.12	0.10	11,449
Middle education	0.67	0.60	0.70	0.66	11,449
High education	0.19	0.23	0.18	0.24	11,449
N at entry search tree	1,611	1,515	2,278	17,586	22,990

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter.

the Netherlands, while smaller groups are from the UK and Belgium. Table 1 provides the descriptive statistics concerning the personal characteristics of respondents. Note that the survey question concerning employment status is asked preceding the search tree, and that age and education are asked in pages following the search tree. Table 1 shows that between 4 and 11% of the respondents are unemployed, mean age varies around 34 years, about one fifth is highly educated and one sixth has a low level of education.

4. Findings

4.1. Explaining Dropout Rates During Search Tree Completion

What explains the dropout rate in the occupation search tree? Table 2 shows that the dropout rates in the 1st step of the search tree across the four country/language

Table 2. Percentages dropout in the three steps of the occupation search tree and percentages employees, self-employed and unemployed, by country/language combination

	UK	Belgium (French)	Belgium (Dutch)	Netherlands
N at start survey	1,808	1,720	2,473	18,810
Dropout page 2	5.8%	7.3%	4.8%	4.6%
Dropout page 3	5.1%	4.6%	3.1%	1.9%
Dropout page 4 – occ Step 1	9.2%	10.5%	10.4%	14.0%
Dropout page 5 – occ Step 2	3.9%	2.4%	3.0%	2.3%
Dropout page 6 – occ Step 3	6.0%	3.0%	4.1%	4.2%
Dropout page 7 till end survey	38.1%	45.9%	37.2%	47.8%
Reached end survey	31.9%	26.2%	37.5%	25.2%
Total	100%	100%	100%	100%

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter (N = 24,811 observations at start of survey).

combinations vary between 9 and 14% and in the 2nd and in the 3rd step between 2 and 6%. Hence, almost one in five respondents drop out during search tree completion and more than half of them do so in the 1st step. The table also indicates that the search tree causes approximately one third of total survey dropout. In Hypothesis 1 it is assumed that the dropout rate is dependent on the number of characters read in the most efficient search path and on respondents' interest in the occupation question. Table 3 shows that the number of characters read in the three steps ranges between a minimum of 62-72 and a maximum of 2,543-3,215 in the four combinations.

Binary logistic regression analysis is used to investigate dropout probabilities in each step of the search tree (Table 4). In Step 1 of the search tree, being employed or self-employed lowers the odds ratio of the dropout probability substantially with 88% and 90%, respectively, compared to the reference group of unemployed. In Step 2 and 3 no significant employment status effects are noticed. Hence the dropout of the unemployed respondents occurs in the 1st step of the search tree. By definition the number of characters read in the 1st step is not available and thus not investigated. In the 2nd step no effect of the number of characters read on the dropout is identified, but in the 3rd step a substantial effect is found. The number of characters in the 1st and 2nd step increases the odds ratio of the dropout probability with 0.1% and 0.2% respectively for each character read. So, for example, if the text string has 100 additional characters the dropout rate at Step 3 increases

Table 3. Descriptive statistics of the number of characters read in Step 1, Step 2 and Step 3 of the search tree, by country/language combination. IQR = Inter Quartile Range, SD = Standard Deviation.

	N	Min.	Max.	Median	IQR	Mean	SD
UK							
# characters read in Step 1	1,415	41	742	408	371	416.6	195.9
# characters read in Step 2	1,334	4	307	50	65	65.4	56.3
# characters read in Step 3	1,215	8	1760	133	195	201.9	235.8
# characters read in Step1+2+3	1,215	72	2543	644	343	676.8	316.5
Belgium (French)							
# characters read in Step 1	1,297	43	824	309	434	366.8	239.8
# characters read in Step 2	1,246	4	305	82	96	100.7	68.9
# characters read in Step 3	1,189	6	2378	154	238	241.8	260.6
# characters read in Step1+2+3	1,189	62	2705	622	471	708.4	350.9
Belgium (Dutch)							
# characters read in Step 1	1,968	46	826	300	435	355.0	241.6
# characters read in Step 2	1,883	6	283	79	94	89.1	61.4
# characters read in Step 3	1,768	6	2411	151	230	236.1	279.9
# characters read in Step1+2+3	1,768	67	3196	599	510	675.2	375.1
Netherlands							
# characters read in Step 1	14,846	46	839	313	500	379.0	258.8
# characters read in Step 2	14,363	6	283	73	102	88.6	65.4
# characters read in Step 3	13,564	6	2456	153	214	227.4	253.9
# characters read in Step1+2+3	13,564	63	3215	625	505	690.0	362.1

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter

Table 4. Effect of employment status and number of characters in the search tree on the probability of dropping out during search tree completion (0=no dropout, 1=dropout)

	Step 1 Odds ratio	Step 2 Odds ratio	Step 3 Odds ratio
# characters in Step 1 (41–839)		1.000	1.001***
# characters in Step 2 (4–307)			1.002***
Employee[1]	.123***	1.054	1.137
Self-employed[1]	.097***	.790	1.193
Country UK[2]	.991	1.345	1.524***
Country Belgium (French)[2]	1.065	.845	.724
Country Netherlands[2]	1.229*	.763	1.004
Constant	.792*	.032***	.030***
−2 Log likelihood	16805.52	5401.84	7771.6
N	22,990	19,524	18,824

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter
Reference categories: [1] Unemployed individuals; [2] Country Belgium (Dutch)
Significance levels: *** $p<.001$, ** $p<.005$; * $p<.010$

10 plus 20%. Country controls have been included, but Table 4 reveals that country hardly influences dropout, except for the UK in Step 3. In conclusion, these results confirm Hypothesis 1. The dropout rates in Step 1 of the search tree are influenced by the respondent's interest and in Step 3 they are affected by the search tree item length.

4.2. Explaining Response Time During Search Tree Completion

Hypothesis 2 asks whether the response time in each step of the search tree is related to the search tree item length or to the respondent's valid self-identification, dropout in the next step and respondent's characteristics. To test Hypothesis 2, the response times for each completed step in the search tree have been derived from the server-side time stamps. Unfortunately, no time stamps are available for the last question before the start of the search tree, hence no response time could be computed for Step 1. The response times are measured in rounded seconds with a minimum of one second. Because response times are skewed, the values have been normalised by taking their natural logs, following discussions by Fazio (1990). Extreme outliers have been deleted by removing the 0.1% values in the long upper tail of the distribution. Table 5 shows that for the four country/language combinations, the median response times are between 10 and 13 seconds for the 2nd step and between 13 and 16 seconds for the 3rd step.

To measure respondent's valid self-identification a 'wrong-match' indicator was developed, based on the OEQ on survey page 7 asking if respondents want to add additional information about the occupational title ticked in the search tree. In total, 4,020 respondents have keyed in relevant text in the OEQ (22.6% of the 17,782 who completed the 3rd step of the search tree). Relevant text is defined as text that includes at least two letters and is not a 'no' response to the question. Particularly in Belgium, this percentage is relatively high (29.6% for BE(French) and 50.4% for BE(Dutch)), whereas it is almost equal for the Netherlands and the United Kingdom (18.9% and 16.7% respectively). The

Table 5. Descriptive statistics of the response time in seconds for Step 2 and Step 3 in the search tree

	N	Minimum	Maximum	Median	(IQR)	Mean	(SD)
UK							
Response time Step 2	1,330	1	210	10	10	15.2	16.6
Response time Step 3	1,211	1	269	13	14	18.3	19.6
Belgium (French)							
Response time Step 2	1,242	1	204	13	11	17.4	18.2
Response time Step 3	1,179	1	206	16	16	20.5	19.8
Belgium (Dutch)							
Response time Step 2	1,887	1	187	11	10	16.7	18.2
Response time Step 3	1,762	1	235	14	14	19.6	21.9
Netherlands							
Response time Step 2	14,321	1	223	11	10	16.1	16.6
Response time Step 3	13,510	1	287	14	13	19.0	20.9

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter

author has compared the ticked occupational title and the answers in the text box, resulting in a classification in six categories (Table 6). The category ADDITIONAL includes either extended task descriptions or refers to composite jobs. An example is: 'I am a secretary with HR tasks'. Most text items fall into this category, demonstrating that the occupational boundaries are not as distinct as the search tree and the occupational classification assume. This problem could be solved by facilitating a second choice in the search tree. An example of 50% MATCH is when the ticked title is 'civil servant in a municipality' and the text box states that the respondent has a clerical job. An example of IRRELEVANT is 'I like my job but not my boss'. The category GENERAL is used particularly in Belgium, where respondents refer to the distinction between blue and white collar workers, which is relevant in this country. The category WRONG reveals that the text includes another occupation than the one ticked in the search tree, thus a wrong match between the search tree data and the text question. The WRONG responses are not equally distributed over the occupational titles in the search tree. The four titles with the most frequent WRONG answers are 'Craft or related worker, all other', 'Paramedical practitioner, all other', 'Process controller, all other', and 'Sales representative'. Similar to the 'not elsewhere classified' occupations in office coding, in search trees in web surveys the category 'all other' fills easily. For these four occupations, the search paths in the occupation database need revision. In total 7.5% of OEQ respondents or 1.7% of respondents with a valid response on the search tree could not identify their occupational title.

OLS regression analysis has been applied to investigate the response time, measured in log seconds, in Step 2 and in Step 3 (Table 7). Model 1 estimates response time without education and age and Model 2 does so with education and age. Two models are used because the number of observations is higher for employment status (asked on page 1) than for education and age (page 10) due to dropout during survey completion. The results confirm Hypothesis 2. The response times in Steps 2 and 3 are indeed influenced by the search tree item length: response times are significantly longer when more characters have

Table 6. The categories and frequencies of responses to the OEQ question on occupation compared to the ticked occupation

Match category	Explanation	% of valid OEQ after search tree	% of valid response in search tree
PERFECT	Text and ticked occupational title are similar	3.6	0.81
ADDITIONAL	Text provides additional information to ticked occupational title	69.7	15.8
50% MATCH	Text indicates that ticked occupational title is not wrong, but the search tree has better alternatives	13.5	3.1
IRRELEVANT	Text is irrelevant given ticked occupational title	4.8	1.1
GENERAL	Text refers to an aggregated occupational title compared to ticked occupational title	0.9	0.2
WRONG	Text indicates that ticked occupational title is wrong	7.5	1.7
		100	22.6
		(N = 4,020)	(N = 17,782)

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter

to be read. For every additional character read in Step 2, the response time in Step 2 is 0.2% larger in both models. For every additional character read in Step 3, the response time in Step 3 is 0.1% larger in both models. This effect is hardly noticeable for the number of characters read in Step 1 affecting the response time in Step 2 and in Step 3, and it is not noticeable for the number of characters read in Step 2 affecting the response time in Step 3. The results also show that for the respondents who drop out in Step 3 the response time in Step 2 is 17% higher, which is in accordance to the findings on the dropout probabilities in the previous section. Finally the results show that the 'wrong-match' respondents need substantially more time in both Step 2 and Step 3, as their response time is 22% and 27% larger respectively (Model 1).

In contrast to expectation, Table 7 shows that the response time is not influenced by the respondent's interest in the occupation survey question: no significant difference between the employed, self-employed and unemployed is found in any of the four models. It makes sense that the unemployed are more likely to drop out in the 1st step, but if they do not, there are no obvious reasons for why they would need more response time. As expected, response times are significantly influenced by respondents' age and educational characteristics: the less educated need more time in Step 3, the highly educated need less time in Steps 2 and 3, and for every additional year of age respondents need 7% more time in Steps 2 and 3. The analyses are controlled for country, revealing that only respondents in Belgium(French) need more time to complete Step 3.

Table 7. Effect of respondent and survey characteristics on the log response time of Step 2 and Step 3 in the occupation search tree (unstandardised coefficients of OLS regressions, standard errors are in parentheses)

	Log response time Step 2		Log response time Step 3	
	Model 1	Model 2	Model 1	Model 2
(Constant)	2.411***	2.197***	2.448***	2.195***
	(−.028)	(−.041)	(−.03)	(−.042)
# characters in Step 1 (41–839)	.000***	.000***	.000	.000**
	(.000)	(.000)	(.000)	(.000)
# characters in Step 2 (4–307)	.002***	.002***	.000	.000
	(.000)	(.000)	(.000)	(.000)
# characters in Step 3 (6–2456)			.001***	.001***
			(.000)	(.000)
Dropout in Step 3	.165***	−.220		
	(−.023)	(−.478)		
Wrong match according to OEQ	.221***	.169***	.275***	.252***
	(−.039)	(−.047)	(−.041)	(−.048)
Employee[1]	−.022	−.020	−.030	−.004
	(−.021)	(−.027)	(−.022)	(−.027)
Self-employed[1]	.006	−.014	−.070	−.040
	(−.028)	(−.036)	(−.03)	(−.037)
Education low[2]		.052		.129***
		(−.022)		(−.022)
Education high[2]		−.125***		−.136***
		(−.016)		(−.016)
Age (10–80)		.007***		.007***
		(−.001)		(−.001)
Country UK[1]	−.053	−.049	−.013	−.010
	(−.024)	(−.031)	(−.026)	(−.032)

Table 7. Continued

	Log response time Step 2		Log response time Step 3	
	Model 1	Model 2	Model 1	Model 2
Country Belgium (French)[3]	.047	.062	.085**	.125***
	(−.025)	(−.032)	(−.027)	(−.033)
Country Netherlands[3]	−.007	.006	.015	.018
	(−.017)	(−.021)	(−.018)	(−.021)
Adj R Sq	0.049	0.07	0.108	0.133
N	18,717	10,696	17,644	10,692

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter
Reference categories: [1] Unemployed individuals; [2] Education middle; [3] Country Belgium (Dutch)
Significance levels: *** $p<.001$, ** $p<.005$; * $p<.010$

These results confirm most of Hypothesis 2. The response time increases with search tree item length, with next-step dropout, with invalid self-identification, with higher age and lower education, but it is not affected by employment status.

4.3. *Explaining Survey Dropout from Search Tree Response Time*

Six to seven in ten respondents do not complete the survey (Table 2). Hypothesis 3 assumes that the total survey dropout is influenced by the search tree response time and valid self-identification, as well as by the respondent's interest, age and education. Table 8 holds the results of a binary logistic regression analysis on the survey dropout for the respondents who have completed at least the search tree on page 6 (Model 1) and the education question on page 10 (Model 2).

Both models reveal that the time-consuming search tree does not influence total survey dropout. Obviously, once the search tree hurdle is taken, its response time does not affect total dropout. Model 2 reveals that being a 'wrong-match' respondent does not influence the survey dropout, but having keyed in relevant text in the OEQ after the search tree does decrease the odds ratio by 35%. Furthermore, Model 2 shows that the odds ratios for the dropout probability increase by 54% for the less and decrease by 13% for the highly educated compared to those with a middling educational level. Neither interest (employment status) nor age affect survey dropout. In both models the country dummies are significant, showing that the odds ratios increase for respondents from Belgium(French) and from the Netherlands compared to those from Belgium(Dutch). This shows the need for explanations beyond this article. In summary, Hypothesis 3 is not

Table 8. Effect of response time on the probability of dropping out at the end of the questionnaire (0=no dropout, 1=dropout) after search tree completion (Model 1) and after completion of the education question (Model 2)

	Model 1 Odds ratio	Model 2 Odds ratio
Response time Step 2 (log)	1.000	1.010
Response time Step 3 (log)	1.037	1.051
Wrong match according to OEQ (0,1)		1.101
Responded to OEQ (0,1)		.645***
Employee[2]	.909	.857
Self-employed[2]	1.203	1.258
Education low[3]		1.540***
Education high[3]		.866***
Age (10–80)		1.001
Country UK[1]	1.196	.911
Country Belgium (French)[1]	1.869***	1.489***
Country Netherlands[1]	1.821***	1.513***
Constant	1.042	.582***
−2 Log likelihood	22829.837	14345.539
N	17,610	10,676

Source: *WageIndicator* survey, Belgium, UK, Netherlands, 2010 second quarter
Reference categories: [1] Country Belgium (Dutch); [2] Unemployed individuals; [3] Education middle
Significance levels: *** $p<.001$, ** $p<.005$; * $p<.010$

confirmed with respect to the effects of the search tree response times, the valid self-identification, interest and age. The respondent's education and country do influence survey dropout.

5. Conclusions and Discussion

Occupation is a key variable in socioeconomic research. Most surveys employ an open-ended question with field or office coding, but problems are associated with this method. The response to the OEQ includes very detailed and very crude occupational titles, and hence the level of aggregation in the occupational classification may vary across respondents. Unidentifiable or ambiguous responses cannot be coded, and this problem is particularly associated with CAPI and CAWI survey modes. Coding is an inexact process within countries and particularly across countries, hampering cross-country analyses. Finally, coding efforts are costly, particularly in case of large-scale multi-country surveys. For these reasons the continuous 70-country *WageIndicator* web survey with large numbers of respondents does not apply an OEQ, but uses a closed format question for which a three-step search tree and a multilingual database with 1,700 occupational titles has been developed, assuming that respondents are able to self-classify their job title into these occupational titles. Occupation search trees are hardly ever used in web surveys and no information is available with respect to the performance of this survey tool. Search trees are assumed to be cognitively demanding and time-consuming. To evaluate the data quality of an occupation search tree, this article explores the dropout rates, the response times and the validity of the ticked occupation from the 2010 second quarter *WageIndicator* web survey in the United Kingdom, Belgium(Dutch), Belgium(French) and the Netherlands.

The first conclusion is that the dropout rates during the occupation search are high, that is, approximately 20%, which is about one third of total survey dropout. The study shows that in the 1st step of the search tree the dropout probability increases substantially for the unemployed respondents, who may judge the occupation question as not adequate for their situation and hence display lower interest in completing the survey. The high dropout rates in Step 1 may also reflect a cognitively demanding task for respondents, who are trying to fit their job titles into highly aggregated categories. The study also shows that the dropout rates in the search tree are influenced by search tree item length, because the number of characters in the 1st and 2nd step increases the odds ratio of dropout in Step 3 with 0.1% and 0.2% respectively for each character read.

The second conclusion is that the median response times are between 10 and 13 seconds for the 2nd step and between 13 and 16 seconds for the 3rd step of the search tree (no data is available for the 1st step). Response times are largest in Belgium(French) and smallest in the UK. The logistic analysis show that the response time, measured in log seconds, is affected by search tree item length, in Step 2 with 0.2% and in Step 3 with 0.1% for every character read in the respective step. Respondents who drop out in Step 3 need 17% more response time for Step 2 and respondents who ticked an invalid occupational title in the search tree need 22% more time in Step 2 and 27% in Step 3. In line with earlier research, the response times are higher for the less-educated and older respondents and lower for the highly educated.

The third conclusion is that the validity of the occupation data is rather good. For this purpose, an open-ended question after the search tree asking for additional information about the ticked occupation was compared with the ticked occupation. More than one fifth of the respondents who completed the search tree used this OEQ. Only 7.5% of these comments indicated that the respondents had not been able to self-identify their occupational title. If all respondents who were unable to identify their occupation had completed the OEQ, the percentage of invalid answers would have been 1.7. Thus the invalid answers are between 1.7 and 7.5% of the respondents completing the search tree. The OEQ reveals another problem. More than two thirds of the OEQ comments refer to additional tasks in the job, suggesting that the occupational boundaries are not as distinct as the search tree and the occupational classification assume. If generalised to all respondents who completed the search tree, approximately 15% of them would have a composite occupation with broader occupational boundaries than suggested in the occupational title in the search tree.

Taking into account its substantial dropout rates and response times, a 3-page search tree apparently is not an optimal response format for the occupation question in web surveys, though recent techniques may in part solve the problems described. First, single page filtering instead of a 3-page search tree most likely will reduce both dropout and response time. Second, the use of text string matching (TSM) may do so even more. In an experiment offering 48 possible values, Funke and Reips (2007) show that these dynamic lists are feasible and that the response time is lower compared to radio buttons. Similarly to search engines, TSM uses dynamic lists with either auto-completion or suggestions for self-identification of occupation, drawing from the WISCO database of occupations. In combination with a single page search tree, TSM may lead to better quality data. If extended with a 'suggest new entry' box, the number of occupational titles in the WISCO database could grow. If made accessible through an Application Programming Interface (API), the tool could offer the research community a sound instrument for the occupation question in web surveys. Increasing use of multi-country web surveys may favour the use of a closed instead of an open survey question. The problem of the composite occupations could be solved by allowing respondents to tick more than one occupation in the search tree.

There are several limitations to this study. The data from only a limited set of countries has been investigated; thus the findings cannot be generalised to all industrialised countries, particularly because some country effects were found. The data has drawbacks. For instance, the time stamps of the question before the search tree were not available and no information was provided about the respondents' back-and-forth clicking in the search tree. Furthermore, the study did not investigate the validity of the occupation variable through a multitrait-multimethod approach. Finally, the results are based on a volunteer web survey, and a detailed comparison of the 2009 Netherlands *WageIndicator* data with a representative reference web survey has demonstrated that there is obviously still a difficulty in quantifying the quality of a nonprobability survey (Steinmetz et al. 2014). Hence the research results presented here should be considered explorative rather than representative. However, given the increasing popularity of web surveys and the urgent need to collect high quality occupation data in these surveys, particularly in multi-country surveys, the study

definitely improves insights into the do's and don'ts of the occupation question for web surveys.

6. References

Armstrong, J.S., Denniston Jr., W.B., and Gordon, M.M. (1975). The Use of the Decomposition Principle in Making Judgments. Organizational Behavior and Human Performance, 14, 257–263. DOI: http://www.dx.doi.org/10.1016/0030-5073(75)90028-8

Blumenstiel, J.E., Roßmann, J., and Steinbrecher, M. (2010). Breakoff in Web Surveys of the German Longitudinal Election Study (GLES). Paper presented at the General Online Research Conference (GOR) 2010. Available at: http://www.websm.org/db/12/13906/Bibliography/Breakoff_in_Web_Surveys_of_the_German_Longitudinal_Election_Study_GLES/ (accessed November 1, 2013).

Callegaro, M., Yang, Y., Villar, A. (2011) Should We Use the Progress Bar in Online Surveys? A Meta-Analysis of Experiments Manipulating Progress Indicators. Mannheim: Paper presented at the General Online Research Conference (GOR). Available at: http://conftool.gor.de/conftool11/index.php?page=browseSessions&presentations=hide&form_session=4 (accessed November 1, 2013).

Conrad, F.G. and Couper, M.P. (2008) Classifying Open Occupation Descriptions in the Current Population Survey. Ann Arbor, MI: Paper presented at the conference on Optimal Coding of Open-Ended Survey Data, University of Michigan.). Available at: http://www.electionstudies.org/conferences/2008Methods/ConradCouper.pdf (accessed November 1, 2013).

De Vries, J. and Ganzeboom, H.B.G. (2008). Hoe meet ik beroep? Open en gesloten vragen naar beroep toegepast in statusverwervingsonderzoek. Mens & Maatschappij, 83, 70–95.

Elias, P. (1997). Occupational Classification (ISCO-88). Concepts, Methods, Reliability, Validity and Cross-National Comparability. Paris: OECD Labour Market and Social Policy Occasional Papers, No. 20. DOI: http://www.dx.doi.org/10.1787/304441717388

Elias, P. and Birch, M. (2006). The Review of ISCO-88: A European Perspective. Warwick: University of Warwick, Working paper Institute for Employment Research.

Elias, P. and McKnight, A. (2001). Skill Measurement in Official Statistics: Recent Developments in the UK and the Rest of Europe. Oxford Economic Papers, 3, 508–540. DOI: http://www.dx.doi.org/10.1093/oep/53.3.508

European Union (2009). Commission Regulation (EC) No. 1022/ of 29 October amending Regulations (EC) No. 1738/2005, (EC) No. 698/2006 and (EC) No. 377/2008 as regards the International Standard Classification of Occupations (ISCO). Official Journal of the European Union, L 283/3, 30 October. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:283:0003:0004:EN:PDF (accessed November 1, 2013).

Eurostat (2009). Task Force on the Quality of the Labour Force Survey. Final report. Luxembourg: Publications Office of the European Union. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-09-020/EN/KS-RA-09-020-EN.PDF (accessed November 1, 2013).

Fazio, R.H. (1990). A Practical Guide to the Use of Response Latency in Social Psychological Research. In Research methods in personality and social research, C. Hendrick and M.S. Clark (eds). Newbury: Sage.

Funke, F. and Reips, U.D. (2007). Dynamics Form: Online Surveys 2.0. Leipzig: Presentation at the General Online Research conference, 26-28 March 2007. Available at: http://frederikfunke.net/papers/pdf/Funke&Reips(2007)Slides_Dynamic_Forms.pdf (accessed November 1, 2013).

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. Journal of Official Statistics, 22, 313–328.

Ganzeboom, H.B.G. (2010). Occupation Coding using ISCO-08. Training session for PIAAC, Bologna, January 20th. Amsterdam: Free University Amsterdam. Available at: http://www.harryganzeboom.nl/Pdf/2010-Ganzeboom-ISCO08-PIAAC-Bologna-(presentation).pdf (accessed November 1, 2013).

Greenwood, A.M. (2004). Updating the International Standard Classification of Occupations, ISCO-08. Geneva: ILO Bureau of Statistics. Available at: http://millenniumindicators.un.org/unsd/class/intercop/training/escwa04/escwa04-9.PDF (accessed November 1, 2013).

Heerwegh, D. and Loosveldt, G. (2006). Survey Length Statements, Progress Indicators, and Survey Sponsor Logos. Journal of Official Statistics, 22, 191–210.

Hunter, D. (2009). ISCO-08 Draft definitions. Geneva: ILO. Available at: http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm (accessed November 1, 2013).

Kaczmirek, L. (2009). Human-Survey Interaction: Usability and Nonresponse in Online Surveys. Neue Schriften zur Online-Forschung, 6. Cologne: Halem.

Lenzner, T., Kaczmirek, L., and Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. Applied Cognitive Psychology, 24, 1003–1020. DOI: http://www.dx.doi.org/10.1002/acp.1602

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. Public Opinion Quarterly, 72, 914–934. DOI: http://www.dx.doi.org/10.1093/poq/nfn050

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. Experimental Psychology, 49, 243–256, DOI: http://www.dx.doi.org/10.1027/1618-3169.49.4.243

Steinmetz, S., Tijdens, K.G., Bianchi, A., and Biffignandi, S. (2014). Improving Web Survey Quality – Potentials and Constraints of Propensity Score Weighting. Online panel research: A Data Quality Perspective, ed. M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavraskas. Chichester: Wiley (forthcoming).

Tippins, N.T., and Hilton, M.L. (Eds) (2010). A Database for a Changing Economy: Review of the Occupational Information Network (O*NET). Panel to Review the Occupational Information Network (O*NET). Washington (DC): The National Academies Press. Available at: http://www.nap.edu/catalog/12814.html (accessed November 1, 2013).

Tijdens, K.G., Van Zijl, S., Hughie-Williams, M., Van Klaveren, M., and Steinmetz, S. (2010). Codebook and Explanatory Note on the WageIndicator Dataset, a Worldwide, Continuous, Multilingual Web-Survey on Work and Wages with Paper Supplements. Amsterdam: University of Amsterdam, AIAS Working Paper 102. Available at:

http://www.uva-aias.net/uploaded_files/publications/WP102-Tijdens,vanZijl,Hughie-Williams,vKlaveren,Steinmetz.pdf (accessed November 1, 2013).

Tijdens, K.G. (2010). Measuring Occupations in Web-Surveys, the Database. Amsterdam: University of Amsterdam, AIAS Working Paper 86. Available at: http://www.uva-aias.net/uploaded_files/publications/WP86-Tijdens.pdf (accessed November 1, 2013).

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

Van der Laan, P., and Van Nunspeet, W. (2009). Modernising Household Surveys in the Netherlands: Design, Efficiency Gains and Perspectives. Paper prepared for the European Directors of Social Statistics Seminar "Re-engineering of Social Statistics: A Perspective", Luxembourg, 25 September 2009. The Hague/Heerlen: Statistics Netherlands, Discussion paper 09044. Available at: http://www.cbs.nl/NR/rdonlyres/0D30D23B-FE40-4570-B41A-E9B2CADF01DB/0/200944x10pub.pdf (accessed November 1, 2013).

Yan, T. and Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. Applied Cognitive Psychology, 22, 51–68. DOI: http://www.dx.doi.org/ 10.1002/acp.1331

Received March 2011
Revised October 2012
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 45–62, http://dx.doi.org/10.2478/jos-2014-0003

Can I Just Check. . . ? Effects of Edit Check Questions on Measurement Error and Survey Estimates

Peter Lugtig[1] *and Annette Jäckle*[2]

Household income is difficult to measure, since it requires the collection of information about all potential income sources for each member of a household. We assess the effects of two types of edit check questions on measurement error and survey estimates: within-wave edit checks use responses to questions earlier in the same interview to query apparent inconsistencies in responses; dependent interviewing uses responses from prior interviews to query apparent inconsistencies over time. We use data from three waves of the British Household Panel Survey (BHPS) to assess the effects of edit checks on estimates, and data from an experimental study carried out in the context of the BHPS, where survey responses were linked to individual administrative records, to assess the effects on measurement error. The findings suggest that interviewing methods without edit checks underestimate non-labour household income in the lower tail of the income distribution. The effects on estimates derived from total household income, such as poverty rates or transition rates into and out of poverty, are small.

Key words: Dependent interviewing; validation study; record linkage; British household panel survey; income; poverty.

1. Introduction

Household income is a key measure of social welfare and as such important for policy analyses. Some surveys, such as the European Social Survey, ask one household member a single question about their income: *"Using this card, please tell me which letter describes your household's total income, after tax and compulsory deductions, from all sources? If you don't know the exact figure, please give an estimate."* Surveys for which income is a key outcome measure more commonly ask a host of questions about each potential source of income, including questions about receipt status, timing of receipt and amounts received. Total income has to be computed from these questions and aggregated over all income sources and all household members. In both cases, reporting on household income is a difficult task for respondents. As a result, household income is likely to be measured with error and estimates derived from it, such as poverty rates or income dynamics over time, may be biased.

In this article we assess the effects of edit check questions, which are incorporated into the questionnaire to detect and correct potential reporting errors, on estimates derived

[1] Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC, Utrecht, the Netherlands, and University of Essex. Email: p.lugtig@uu.nl
[2] University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK. Email: aejack@essex.ac.uk
Acknowledgments: This work was supported by the European Centre for Analysis in the Social Sciences (ECASS), which funded a visit of the first author to the University of Essex. The second author gratefully acknowledges funding from the ESRC (RES-000-22-2323). Data collection for the experimental validation study was funded by the ESRC Research Methods Programme (H333250031). We are grateful to Peter Lynn, Stephen P. Jenkins and Gerty Lensvelt-Mulders for comments on earlier versions of the article.

from detailed questions about household income. We examine the effects of both within-wave and cross-wave edit checks in the measurement of non-labour household income. *Within-wave edit checks* use information collected earlier in the same interview to check the consistency of answers. For example, respondents can be queried about sources they have not reported, but for which they are likely to be eligible, judging from responses given earlier in the interview (Pennell 1993). *Cross-wave edit checks* are specific to longitudinal surveys. They use information provided in previous interviews to check the longitudinal consistency of responses. For example, respondents can be queried about sources they have reported in the past, but not in the current interview (see Jäckle 2009; Mathiowetz and McGonagle 2000). Cross-wave edit checks are typically referred to as 'dependent interviewing' (DI) and we follow this convention.

The key question examined here is to what extent edit checks affect estimates of household income and poverty. Previous studies evaluating the effects of DI have mainly focused on measurement error in receipt status for individual income sources. These studies have shown that some non-labour income sources are considerably underreported and that DI improves reporting for non-labour income (Lynn et al. 2012). Other studies have examined measurement error in the timing of receipt and shown that DI reduces errors in monthly transition rates (Moore et al. 2009) and spell durations (Jäckle 2008). The effects on monetary amounts have not been examined to our knowledge. Neither have the effects on estimates related to total (household) income. Although the reduction of error in individual survey questions can be substantial, it is not a priori clear what effect this methodological improvement has on estimates that are derived from a series of detailed questions about all components of household income.

We contribute to this literature by examining to what extent edit checks affect estimates of household income, poverty rates and transitions into and out of poverty. For this purpose we use three waves of the British Household Panel Survey (BHPS), in which both within-wave edit checks and DI are used in a quasi-experimental way for the collection of non-labour income data. The analyses of the BHPS data illustrate to what extent edit checks affect estimates derived from income data. The BHPS data however do not allow any conclusions about the effects of edit checks on measurement error and resulting biases in estimates. We therefore complement these analyses using data from an experimental study carried out in the context of the BHPS, which linked survey responses to individual administrative records.

The results suggest that traditional methods of interviewing that do not use edit checks for non-labour income sources underestimate household income in the lower tail of the income distribution. Estimated poverty status and poverty transitions however hardly change. The changes in estimates appear to reflect a reduction in measurement error in the reporting and duration of receipt, thus reflecting an improvement of data accuracy.

2. Data

2.1. The British Household Panel Survey (BHPS)

The BHPS is a panel survey of the UK population that started in 1991 with a clustered and stratified address-based sample of 5,500 households. All household members aged 16+

are interviewed annually and followed as long as they remain in the UK. The individual response rates, conditional on response in the prior wave, are around 94% (RR1 – AAPOR 2011) in the waves we used for our analyses (Waves 15-17). All in all, 49.6% of the original sample members with an interview at Wave 1 completed an interview in Waves 15-17 (Taylor et al. 2009). Proxy interviews were held with about 1% of sample members in each wave and treated as missing data in our analyses.

Income data are collected in two sections of the questionnaire: one on labour earnings and another on non-labour income (including state cash transfers, private pensions, private transfers and investment income). Edit checks are only used for non-labour income and we therefore focus on those questions. We do however include data about labour income to assess the effects of edit checks on derived measures of total household income.

In the original version, respondents are shown a series of four showcards, listing 34 potential income sources, and asked which of these they have received during the reference period. Fieldwork takes place between September and January each year. Respondents are asked to report about the period since the start of fieldwork in the previous year. This means that, depending on the month in which a respondent is interviewed, the recall period covers between 12 and 16 months: *"Please look at this card and tell me if, since September 1st <previous calendar year>, you have received any of the types of income or payments shown, either just yourself or jointly?"* For each income source reported, respondents are then asked a series of follow-up questions about the timing and amounts of receipt: *"And for which months since September 1st <previous calendar year> have you received <source>?"*, *"How much was the last payment of <source> you received?"*, and *"What period did that cover?"*

From 2005 the BHPS added within-wave edit checks for those cash transfers, for which questions earlier in the same interview predict eligibility: Pension Credit, Disability Benefits, Income Support, Jobseeker's Allowance, Child Benefit and Housing Benefit. For example, respondents above the state retirement age who have not reported a state pension are asked *"Can I just check, do you currently receive the State Retirement Pension?"*

From 2006 onwards, reactive dependent interviewing (RDI) was added for all non-labour income sources (listed in Subsection 2.4). Respondents are first asked the original question. For any income sources reported in the previous but not the current interview, they are asked a follow-up question: *"Can I just check, according to our records you have in the past received <source>. Have you received <source> at any time since <date of interview>?"* (see Jäckle et al. 2007).

Although the BHPS data are not experimental, the public release file identifies which income sources were reported in response to the initial question, which in response to the within-wave edit check, and which in response to the RDI follow-up question. This enables a quasi-experimental comparison of the effects of the interviewing method on responses and estimates.

2.2. The Experimental Validation Study

The experimental study was carried out using the former European Community Household Panel (ECHP) low-income subsample for Great Britain. This sample was surveyed as part of the BHPS (using the BHPS survey procedures and questionnaires) from 1997 until

funding expired in 2001. In 2003 the sample was interviewed once more for methodological purposes. The experimental survey included a split-ballot experiment comparing independent and dependent interviewing for various sections of the questionnaire. In addition, respondents were asked for permission to link to their records on receipt of 17 different state cash transfer programmes held by the Department for Work and Pensions (the department in charge of administering cash transfers). The transfer programmes included Child Benefit, Housing Benefit, Working Families' Tax Credit, different types of Disability Allowances, Income Support, Jobseeker's Allowance and State Pensions.

The response rate for the experimental survey was 89% (N=1,033, RR1–AAPOR 2011), of which 77% gave consent for the record linkage (Jäckle et al. 2004), of which 74% were successfully linked. A related study by Jenkins et al. (2006) found that households that had reported receipt of means-tested state cash transfers in a previous wave of the survey were more likely to consent to the data linkage, but that household income was not related to consent. The linkage was performed independently five times using deterministic (exact) matching on National Insurance Number (the UK social security number) or sex with two or three out of date of birth, postcode, first line of address, first name, and family name (see Jenkins et al. 2008 for details on the linkage methodology). Results for each respondent were pooled to identify a single match. For 12 of the 14 respondents who were matched to more than one person in the administrative records, the modal match (which matched on at least three of the five criteria) was used as the correct match. The other two cases were inspected visually to determine the correct match. Although some problems with the linkage variables cannot be excluded, Jenkins et al. (2008) suggested that those not linked were probably respondents who had not received state cash transfers during the time frame of interest. The authors estimated that the true non-match rate was about one quarter, since 29% of respondents never reported receiving any of the relevant state cash transfers in any of the annual interviews between 1999 and 2003.

In the experimental survey, three versions of questions on non-labour income components were randomly assigned: independent interviewing (INDI, N=348 respondents of which N=262 consented to the record linkage), reactive dependent interviewing (RDI, N=344 respondents of which N=274 consented) and proactive dependent interviewing (PDI, N=341 respondents of which N=263 consented). With PDI respondents were reminded upfront of each source they had reported in the previous interview, and asked whether they had received the source since. Since the BHPS uses RDI for the income questions, the analyses presented here focus on the comparison of INDI and RDI. The INDI version used the original BHPS question, as described in Subsection 2.1. The RDI version had an added edit check question, again as described in Subsection 2.1.

Respondents in both experimental conditions were asked the same series of follow-up questions, described in Subsection 2.1, about the timing and amounts of each income source. The administrative records contain information about which of the 17 potential state cash transfers listed above each respondent has received, including the exact start and end dates of receipt and weekly amounts received. The data stem from the database system used by the state to administer transfer payments and are generated in the process of payments being made. This means that the administrative data reflect the actual dates and amounts of payments and can therefore be considered high quality (except for Housing Benefit data which are from decentralized databases and less reliable). A few transfer

types included in the survey are not included in the records (Widowed Mother's Allowance, War Disability Pension, Council Tax Benefit). Some cash transfer types (Disability Living Allowance, Child Benefit) are recorded as a single source, while the survey collects separate information about different components (e.g., care component vs. mobility component). For comparability, we derived variables from the experimental survey data that reflect the data structure and definitions of the record data.

2.3. Comparability of the BHPS and Experimental Survey Data

Although the survey data from the experimental study and the BHPS are based on the same design, there are several differences between the surveys which are relevant to our analyses:

(1) Time frames: The BHPS data are from 2005, 2006 and 2007, while the experimental survey data are from 2001 and 2003.
(2) Sample composition: the BHPS is a general population sample, while the experimental survey data overrepresent low-income households and may be affected by selection bias due to non-consent to linkage.
(3) Dependent interviewing method: the BHPS used RDI for all sample members in 2006 and 2007, while the experimental survey used INDI in 2001 and experimentally allocated respondents to a DI treatment in 2003.
(4) Within-wave edit checks: the BHPS used within-wave edit checks for questions on cash transfer receipt in 2005, 2006 and 2007 surveys, while the experimental survey did not use any within-wave edit checks.

These differences between the survey data from the BHPS and the survey data in the experimental validation study mean that it is not clear a priori whether the results from the validation study are likely to apply to the BHPS survey. We report on additional analyses we have carried out to verify the comparability of data from the two surveys in the discussion in Section 4.

2.4. Data Description

For analysis purposes, we group the income sources into four components of non-labour income: State cash transfers, private pensions, other transfers, and investments. This grouping corresponds to the derived income components provided with the BHPS public release file and consists of the following income components:

(1) State cash transfers: four types of national insurance pensions and tax credits, ten types of disability-related cash transfers and tax credits, two types of income support, Housing Benefit, Council Tax Benefit, Jobseeker's Allowance, Child Benefit, Maternity Allowance, Working Families' Tax Credit, Child Tax Credit.
(2) Private pensions: three types of private pensions.
(3) Other transfers: education grants, sickness insurance, maintenance/foster allowance, payments from trade unions/friendly societies, payments from absent family members, other payments.
(4) Investment income: rent from boarders/lodgers, rent from other properties.

Labour income also contributes to household income. We do, however, not examine this component separately, because edit check questions were not used for the collection of labour income data.

3. Results

3.1. Effects of Edit Checks on Survey Estimates

The number of income sources reported in the BHPS is documented in Table 1, for the 2005 survey (Wave 15), 2006 (Wave 16), and 2007 (Wave 17). In Wave 16, for example, respondents reported receipt of a total of 8,170 state cash transfers when asked the original BHPS question. Respondents for whom information collected earlier in the interview suggested that they might be eligible for additional cash transfers were then asked the within-wave edit check, whereupon they reported a further 165 sources. Finally, all respondents were queried about sources they had reported in the previous interview using the RDI edit check question, whereupon they reported a further 615 income sources. As described in Subsection 2.1, the within-wave edit check questions were only used for state cash transfers. For the other types of non-labour income only the RDI edit check was used.

The results suggest that RDI edit checks were more effective at increasing reporting of income sources than within-wave edit checks. Depending on the type of income and the survey year, about 1-2% of total income sources were reported in response to the within-wave edit checks, whereas 5-12% were reported in response to RDI edit check questions.

The sample sizes in the experimental survey and validation data are documented in Table 2. Of the sample allocated to INDI, 262 respondents consented to the linkage. These respondents reported a total of 338 state cash transfers in the 2003 survey, while the administrative records for these respondents, corresponding to the same time period, list 374 cash transfers. The fact that respondents in aggregate reported fewer income sources in the survey than they received according to the records is a first indication of underreporting. In contrast, respondents allocated to RDI reported 401 state cash transfers with 407 recorded in the administrative data, suggesting that RDI improved the aggregate reporting of income sources.

Table 1. Number of income sources reported in the BHPS

		Cash transfers	Pensions	Other transfers	Investment
Wave 15	INDI	8,088	1,717	426	274
	WWEC	117 (1.4%)	–	–	–
Wave 16	INDI	8,170	1,776	515	323
	WWEC	165 (1.8%)	–	–	–
	RDI	615 (6.9%)	121 (6.4%)	55 (9.6%)	39 (10.8%)
Wave 17	INDI	7,895	1,846	501	302
	WWEC	157 (1.8%)	–	–	–
	RDI	506 (5.9%)	94 (4.8%)	49 (8.9%)	42 (12.2%)

Notes: Number of respondents in Wave 15: 8,538; Wave 16: 8,484; Wave 17: 8,322.
INDI: independent interviewing, WWEC: within-wave edit check, RDI: reactive dependent interviewing. Percentages represent the percent of total income sources of a given type reported in response to the WWEC or RDI.

Table 2. Number of income sources in the experimental validation data

Experimental treatment group	Cash transfers	
	in records	in survey
INDI	374	338
INDI+RDI	407	401

Notes: INDI: independent interviewing, RDI: reactive dependent interviewing. Based on 2003 survey.

3.1.1. Effects on the Distribution of Household Income

To examine whether edit checks affect estimates of household income, we use Waves 15 to 17 of the BHPS. Table 3 shows estimates of the equivalised annual household income distribution for the population of Great Britain. The estimates are based on all members of surveyed households, adjusted for differences in household size using the McClements equivalence scale (Taylor et al. 2009) and weighted for nonresponse. The first column indicates the estimated cut-off points between percentiles of the income distribution, including only amounts associated with income sources reported in response to the INDI questions. The income measures based on INDI include imputed values if the receipt status, amount received or dates of receipt are missing for any of the income sources (Taylor et al. 2009). The second column indicates by how much the income percentile changes when income sources reported in response to the within-wave edit checks are included. For Waves 16 and 17, the third column indicates by how much the INDI estimate changes if sources reported both in response to the within-wave edit checks and the RDI follow-up questions are included.

Within-wave edit checks have a considerable effect, increasing estimated income percentiles below median income, for example increasing household income for the fifth percentile by 6% at Wave 16. RDI has an additional effect, increasing the fifth percentile by a further four percentage points to 10%. The effects of RDI and edit checks are largest for people in the lowest percentile, fall monotonically across percentiles, and are zero or close to zero for all percentiles above the median. The effect on median income is small: when sources reported in response to the within-wave edit checks are included, the estimated median increases by less than 0.3% in each of the three waves, and by a further 1% at Waves 16 and 17 when responses to RDI are included.

Edit checks on non-labour income sources therefore increase estimates of household income at the lower end of the income distribution, where non-labour income from cash transfers, pensions, and other transfers represents a major component of total income. For households with higher levels of income, these sources are less important, while non-labour income from investments may contribute a large part of total income. Nonetheless, the edit checks do not have any effect at the upper tail of the income distribution.

3.1.2. Effects on Estimated Poverty Rates

To examine whether edit checks affect estimated poverty rates, we again use Waves 15 to 17 of the BHPS. Replicating the official UK poverty definition, we define the poverty threshold as 60% of median household income: any individual living in a household with

Table 3. Distribution of equivalised annual household income

	Wave 15		Wave 16			Wave 17		
Percentile	INDI (£)	INDI+ WWEC (% Δ)	INDI(£)	INDI+ WWEC (% Δ)	INDI+WWEC+ RDI (% Δ)	INDI (£)	INDI+ WWEC (% Δ)	INDI+WWEC+ RDI (% Δ)
1	1609	49	436	183	210	842	95	137
2	3740	6	2780	31	35	3205	26	32
5	6047	5	6073	6	10	6385	4	7
10	8353	1	8549	2	6	8630	2	6
25	13594	0	13881	0	3	13847	1	4
50	25192	0	25267	0	2	25594	0	1
75	40921	0	41106	0	1	42472	0	1
90	57812	0	58602	0	0	61828	0	0
95	71091	0	73977	0	0	75365	0	0
98	89872	0	93273	0	0	96199	0	0
99	107793	0	109815	1	1	115145	0	0

Notes: Based on all BHPS enumerated household members, Wave 15: 11,700; Wave 16: 11,611; Wave 17: 11,374. Data are weighted for nonresponse and initial sample selection probabilities. INDI: Income derived from independent interviewing questions only, INDI+WWEC: INDI plus within-wave edit checks, INDI+WWEC+RDI: INDI+WWEC+ reactive dependent interviewing. (£): annual equivalised household income. (% Δ): percentage change compared to INDI only.

Table 4. Poverty rates (%)

Wave	Interviewing method	'Poor'	INDI: 'poor' Edit check: 'not poor'	INDI: 'not poor' Edit check: 'poor'
15	INDI	18.6	–	–
	INDI+WWEC	18.5	0.8	0.0
16	INDI	18.9	–	–
	INDI+WWEC	18.8	0.9	0.1
	INDI+WWEC+RDI	18.4	4.2	0.4
17	INDI	18.4	–	–
	INDI+WWEC	18.2	1.2	0.0
	INDI+WWEC+RDI	17.9	3.9	0.3

Notes: Based on all weighted BHPS enumerated household members, Wave 15: 11,700; Wave 16: 11,611; Wave 17: 11,374. INDI: independent interviewing. WWEC: within-wave edit check. RDI: reactive dependent interviewing.

less income is classified as poor. Official poverty statistics use 60% of current income to define the threshold for poverty (Brewer et al. 2009). We use annual income instead, in order to examine the net effects of edit checks on all questions related to household income, including questions about the timing of receipt during the year. In addition, Böheim and Jenkins (2006) show that there are few differences between poverty indicators based on current and annual income.

The results in Table 4 suggest that the edit checks somewhat reduce estimated poverty rates, but the effects are small: both in Waves 16 and 17, adding responses to edit checks and RDI reduced the poverty rate by 0.5 percentage points. Nonetheless, some individuals are classified differently depending on the interviewing method. For example, in Wave 16, 4.2% of individuals are classified as 'poor' based on the INDI questions, and as 'not poor' when the income sources reported in response to the edit checks are added. Similarly, the third column shows that 0.4% of individuals classified as 'not poor' with INDI are classified as 'poor' when information from the edit checks is added. These are probably households whose income is only just above the poverty threshold based on the INDI data, and who did not report any additional income sources in response to the checks or RDI. Since median income, and therefore also the poverty threshold, increases slightly when the edit check responses are included, these respondents slip just below the poverty threshold.

3.1.3. Effects on Estimated Poverty Transitions

To examine whether edit checks affect the longitudinal consistency of poverty classifications across waves, we again use BHPS data. Table 5 shows the transitions in poverty status between Waves 15 and 16, and Waves 16 and 17, based on the INDI data only, adding the within-wave edit check data, and further adding the RDI data. The edit checks have little effect on transition rates in both wave pairs: in the INDI data about 76% of individuals were living in non-poor households in both waves, 13% were poor in both waves, around 5% entered poverty and a further 5% exited poverty from one wave to the next. These estimates are similar when data from the within-wave edit checks and RDI are added. The lack of effects is surprising, since we would have expected RDI to increase the

Table 5. Transition rates into and out of poverty (%)

Wave	Transition type	INDI	INDI+WWEC	INDI+WWEC+RDI
15-16	Persistent non-poor	76.3	76.4	–
	Persistent poor	13.1	13.0	–
	Transition into poverty	5.6	5.7	–
	Transition out of poverty	4.9	5.0	–
16-17	Persistent non-poor	76.3	76.5	77.0
	Persistent poor	13.2	13.2	12.9
	Transition into poverty	5.0	4.8	4.8
	Transition out of poverty	5.5	5.5	5.3

Notes: Based on all weighted BHPS enumerated household members, Wave 15-16:10,278; Wave 16-17: 9,692. INDI: independent interviewing. WWEC: within-wave edit check. RDI: reactive dependent interviewing.

consistency of responses across waves, and by implication to reduce changes in household income and resulting changes in poverty status across waves.

In sum, both within-wave edit checks and RDI increase estimates of household income at the lower end of the distribution, but neither method has much effect on poverty classifications or transitions. The next section examines whether the changes in household income reflect an improvement in data accuracy.

3.2. Effects of RDI and Edit Checks on Measurement Errors

We use the validation data to examine various aspects of measurement error related to the estimates presented in Subsection 3.1. We examine measurement error in receipt status, amounts of income, duration of receipt, and transitions in receipt status between waves. For each of these aspects we examine the extent of measurement error with independent interviewing, and how this changes with RDI edit checks. Note that as described in Subsection 2.1, RDI was only applied to questions about receipt, not to questions about dates or amounts of receipt. The RDI edits on the receipt questions nonetheless affect responses to the amounts and duration questions, since sources that are not reported by default have zero amounts and durations associated with them. The analyses of the experimental validation data are unweighted.

We expect the changes in responses with RDI to reflect a reduction in the various aspects of measurement error, and therefore expect the changes in estimates in Subsection 3.1 to reflect improvements in data accuracy.

3.2.1. Effects of RDI on Measurement Error in Receipt of Income Sources

We first examine the effect of RDI on measurement error in individual reports of non-labour income *receipt*. We compare responses to the experimental survey with individual register data. For each potential income source, we derive indicators of whether or not the source was received at any point during the reference period. Separate indicators are derived for the survey and the record data and used to classify all potential income sources for each respondent: *true negatives* are income sources which were neither received according to the survey, nor according to the records; *true positives* are income sources which were received both according to the survey and the records; *false negatives* are

Table 6. Effect of RDI on measurement error in income receipt reported by individuals

	Sample sizes (N) and row percentages (in brackets)				Error rates (%)	
	True negative	False negative	False positive	True Positive	False negative rate	False positive rate
INDI	3257 (88.8%)	73 (2.0%)	26 (0.7%)	312 (8.5%)	19.0	0.8
RDI	3377 (88.0%)	58 (1.5%)	30 (0.8%)	371 (9.7%)	13.5	0.9

Notes: The sample includes all respondents (INDI=262, RDI=274), multiplied by 14 potential income sources. Columns are defined in the text. INDI: Independent Interviewing, RDI: Reactive Dependent Interviewing.

income sources which were received according to the records, but not reported in the survey; *false positives* are income sources which were not received according to the records, but reported in the survey.

To account for the possibility that respondents may report income sources which are recorded in the name of a different household member in the record data, income sources are counted as 'true positives' if there is a record for the source in the name of another household member. This was the case for 3% of income sources reported by the INDI sample, and 5% of sources reported by the RDI sample. Table 6 indicates the number of potential income sources which are classified as true/false positives/negatives. Assuming that the record data represent the true values, we interpret 'false negatives' as indicators of underreporting, and 'false positives' as overreporting. The last two columns indicate the corresponding error rates: the false negative rate is the number of false negatives as a proportion of sources received according to the records; the false positive rate is the number of false positives, as a proportion of the sources not received according to the records.

The results indicate that the main type of error is underreporting: with INDI 19.0% of sources recorded in the records are not reported in the survey, while overreporting hardly occurs (less than 1%). RDI reduces the false negative rate to 13.5% and does not have any effect on overreporting. The increase in the reporting of income sources with RDI therefore represents a reduction in net measurement error in receipt of non-labour income sources.

3.2.2. Effects of RDI on Measurement Error in the Amounts of Non-Labour Income

Second we test the effects of RDI on measurement error in the *amount* of income, again comparing the survey reports to the individual records. For each source we derive the amount of the last payment during the reference period according to the survey and according to the records. The amounts are standardised to weekly amounts, for comparability with the format in which they are recorded in the administrative data. We then calculate the error in amounts of receipt as the difference between the survey and the record. In the final step, we calculate the mean error over all cash transfers and respondents. The analysis includes sources reported either in the survey, or the records, or both. Housing benefits were excluded in this step as we found large irresolvable consistencies between the records and survey data.

With INDI, weekly non-labour income is underreported by £4.60 on average (95% confidence interval (CI) from −9.1 to −0.2). With RDI the error increases to £5.90

(95% CI from −9.9 to −1.9). This suggests that although RDI reduces underreporting of receipt, it does not help respondents report the amounts received.

3.2.3. Effects of RDI on Measurement Error in Duration of Receipt

Third, to assess the effects of RDI on measurement error in reported *duration* of receipt, we again compare the survey and administrative data. For each income source we calculate the error as the difference between the number of months of receipt according to the survey and the records. The analysis is restricted to receipt between September 1st 2001 and September 1st 2002, for comparability with the BHPS data. The base includes all sources either reported in the survey, or recorded in the administrative data, or both, but excludes true negatives. In the case of overreporting where a record exists in the name of a different household member, the survey duration is compared to the record duration for the other household member. We then calculate the mean error over all income sources and respondents.

With INDI receipt is underreported by 1 month on average (95% CI from −1.4 to −0.4). With RDI the mean error is no longer significantly different from zero (95% C.I. from −0.4 to 0.4). This suggests that RDI reduces measurement errors in reported duration of receipt of cash transfers.

3.2.4. Effects of RDI on Measurement Error in Transitions of Cash Transfer Receipt Across Waves

Fourth, we evaluate whether RDI reduces measurement error in reported *transitions* of receipt across waves. We classify each potential income source for each respondent according to the type of transition between the 2001 survey and the 2003 survey as continued non-receipt, continued receipt, transition off receipt, and transition onto receipt. Each potential income source is classified separately based on the survey data and the record data. We then pool the results for all income sources and compare the transition types derived from the survey and records to identify errors in transition classifications.

Overall, the transition type is misclassified for 4% of potential income sources with both INDI and RDI. Since RDI was only used in the 2003 interview, the interviewing method cannot have affected the wave 2001 status. Therefore Table 7 focuses on errors in the classification of transition types, conditional on the 2001 status being reported correctly in the survey. The rows indicate the respondents' transition statuses (pooled over all potential income sources) according to the records. The columns indicate the percentage of income sources for which the 2003 status was misclassified in the survey, resulting in an error in the transition type.

'Continued non-receipt' is reported well with INDI and not improved with RDI: the error rates are 0.5% with both methods. With RDI more respondents are correctly classified as having 'continued receipt': the error rates are reduced from 11.3% to 3.2%. However, respondents who 'transitioned onto' cash transfer receipt are more likely to be misclassified with RDI: the error rate unexpectedly increases from 20.4% to 38.6%. Further investigation (not shown in the table) suggests that RDI respondents who transitioned onto receipt are more likely to be misclassified as 'continued non-receipt'. This is a surprising finding, since non-receipt in the previous interview does not trigger any RDI questions. 'Transitions off' cash transfer receipt tends to be reported correctly with

Table 7. Effect of RDI on measurement error in transitions onto and off cash transfer receipt, conditional on correct classification in the 2001 survey

	INDI			RDI		
Transition in Records		N mis-classified	% mis-classified	N	N mis-classified	% mis-classified
Continued non-receipt	3230	16	0.5	3348	15	0.5
Transition on	49	10	20.4	57	22	38.6
Continued receipt	284	32	11.3	317	10	3.2
Transition off	16	0	0.0	15	2	13.3

Notes: The sample includes all respondents (INDI=262, RDI=274), multiplied by 14 potential income sources, and excluding sources reported incorrectly in the first interview (INDI=89, RDI=99). INDI: independent interviewing, RDI: reactive dependent interviewing.

INDI, but the number of transitions is very small. With RDI however 13.3% of transitions are misclassified, mostly as 'continued receipt'. This could be due to respondents falsely confirming a receipt status presented to them from the previous interview. A potential cause of the findings for transitions onto and off receipt might be found with the interviewers. With DI designs, they might be more focused on reducing errors in continued receipt than on picking up transitions onto and off receipt (Sala et al. 2009). Since the number of transitions onto and off receipt is small, we would however interpret these results with caution.

In sum, RDI reduces various aspects of measurement error in the reporting of state cash transfers: RDI reduces underreporting of any receipt, of the duration of receipt within one wave, and of continued receipt across waves. RDI does not reduce overall misclassification rates in transitions, although the nature of misclassifications changes.

4. Discussion and Conclusion

4.1. Summary of Results

The motivation for this study was to examine what effect methodological innovations that are expected to reduce measurement error have on substantive estimates. In this case, substantive conclusions are affected by whether or not edit checks are used to collect income data. Methodological studies designed to evaluate the effects of alternative data collection methods on data quality often only examine answers to individual survey questions. Evaluations of the impact on data quality however further need to relate to the actual uses of the survey data. In this spirit, we examine the effects of within-wave edit checks and RDI on derived estimates, and subsequently whether these effects reflect a decrease in measurement error. For this purpose we exploit a unique combination of data sets: we use data from the BHPS, a large-scale panel survey which has implemented within-wave edit checks and RDI for questions on non-labour income components in a quasi-experimental way, and from an experimental validation study based on the BHPS survey design.

We use the experimental study to assess the effects of RDI on different aspects of measurement error, and the BHPS data to assess the effects of RDI and within-wave edit

checks on estimates of household income and poverty. The results suggest that both the within-wave edit checks and RDI increase estimates of total household income in the lower tail of the income distribution. Neither method has much effect on estimated poverty rates or estimated rates of transitions into and out of poverty. The increase in household income reflects an increase in data accuracy: RDI reduces underreporting without affecting overreporting; RDI reduces underreporting of months of receipt and reduces erroneous transitions off income receipt and underreporting of continued receipt across waves.

In our view, the effects of RDI on measurement error are considerable; for example, the underreporting rate is reduced by about 29% compared to independent interviewing. The effects on estimates of household income and poverty are arguably small. This suggests that while within-wave edit checks and RDI may have large effects on measurement error in responses to individual survey questions, the combined effects, in this case over different survey items and different household members, may be small. This conclusion may however be open to interpretation, since a reduction in the estimated poverty rate by a mere 0.5 percentage points affects around 300,000 individuals in the population of Great Britain.

4.2. Limitations

Our study has several limitations. First, we only examined the effects of edit checks on non-labour components of household income. Non-labour income makes up almost 90% of total gross household income for the lowest income quintile. The higher the quintile however, the less important non-labour income is; non-labour income comprises 65% of total income for the second income quintile, 39% for the third, 18% for the fourth and 9% for the highest quintile (Brewer et al. 2009). We cannot draw conclusions about the likely effects of the introduction of edit checks on labour income, because the nature of measurement error in labour and non-labour income is quite different. With non-labour income, measurement error is mainly in the form of underreporting: respondents fail to report receipt of a source, and by default the amount received is set to zero. Overreporting is rare (see Table 6 and Bound et al. 2001). With labour income, respondents are less likely to underreport receipt, as they are unlikely to underreport being in work. Instead, measurement error occurs in the amount of earnings. Errors tend to be negatively correlated with true values and to cancel out across respondents (Bound et al. 2001). The nature of measurement error has implications for the potential effects edit checks can have. For non-labour income sources, we expect edit checks to reduce underreporting and therefore to increase estimates of household income for households for which non-labour income sources represent a major proportion of total income. This corresponds to our findings. For labour income, we would expect edit checks on average to decrease the earnings reported by those with low earnings, and to increase the earnings reported by those with high earnings. As a result, we would expect estimates of household income to change at both ends of the distribution, but because labour income comprises a larger part of total income, we expect the largest changes for the higher quintiles. RDI edit checks have been used to query changes in earnings in the Survey of Labour and Income Dynamics (SLID). Hale and Michaud (1995) reported that 8.3% of respondents reported

earnings that differed by more than ± 10% and were then asked an RDI edit check. Two thirds of respondents confirmed the change as true. The experimental study we use in this article also included a test of RDI edit checks for earnings. In this study, 59% of respondents reported a change in earnings larger than ± 10% and were asked the edit check. All but one confirmed that the change was true (Jäckle 2009). It is not clear from these studies however what the effect of the edit check for amounts of labour income on estimates of household income is.

Second, our validation study is limited in that it contains only data on state cash transfers (and not on the other non-labour income sources for which RDI is used in the BHPS), and in that the study contrasted only INDI and RDI (and did not use within-wave edit checks). To check whether it is reasonable to assume that the findings from the validation study also apply to the BHPS, to the other income sources, and maybe also to the within-wave edit checks, we have carried out some further analyses (see Lugtig and Jäckle 2011). We compared 1) the effects of RDI in the BHPS versus the validation survey, 2) the effects of RDI on reporting of state cash transfers versus other non-labour income sources in the BHPS, and 3) the effects of RDI versus within-wave edit checks in the BHPS. The results suggest that the effects of the questioning method are similar for these three comparisons. We therefore assume that the effects of RDI on measurement error in the validation study are also likely to apply to the BHPS data. We further assume that for those income sources for which we have no validation data, the changes in responses also reflect a reduction in measurement error. Finally we assume that the changes in responses due to the within-wave edit checks also reflect reduced measurement error. As a result, we assume that the changes we find in estimates related to household income represent improvements in data accuracy.

Third, in this article we do not investigate which specific types of income sources are most likely to be misreported, or which types of respondents are most likely to misreport. Lynn et al. (2012) examined the same validation study we use here and reported error rates for the six most common state cash transfer programmes. Among these, underreporting rates were highest for Incapacity Benefit (50%), followed by Tax Credit (29%), Child Benefit (23%), Housing Benefit (17%), Income Support (11%), and lowest for Retirement Pension (0%). Overreporting rates were between 0% and 2% for all sources (rates derived from Table 3). Our own analyses for the same income sources on the BHPS data suggested that the extent of underreporting was generally lower and differences between individual sources were much smaller: underreporting was highest for Housing Benefit (17%), followed by Income Support (8%), Incapacity Benefit (7%), Tax Credit (5%), and lowest for Child Benefit (4%), and Retirement Pension (4%).

Lynn et al. (2006), using the validation data, also examined which types of respondents were most sensitive to RDI edit checks. Respondents who reported income sources only in response to the edit check were less likely to be retired (or born before 1943) or living with a spouse or partner, but more likely to be registered disabled than respondents who reported receipt in response to the independent questions. The authors found no differences by gender, whether in paid work, children in the household, qualifications, general health, duration living at the address, regular car, or mobile phone use (Table 6 in Lynn et al. 2006).

4.3. Future Research

There are a number of issues, regarding both the effects of RDI and within-wave edit checks and the mechanisms through which these methods work, which in our view warrant further attention. Reactive and proactive DI have rarely been compared. The reason why RDI was implemented in the BHPS was that this made it possible to maintain comparability with the previous 15 waves of data collection, in which independent interviewing was used. The responses given to the independent question can still be identified and for comparisons with previous waves the responses to the reactive follow-up can be ignored.

Our ability to compare the effects of within-wave edit checks and RDI were limited by the fact that they were always used in combination, with the edit checks always preceding the RDI checks. With this design, the RDI edits seemed more effective at increasing the reporting of income sources than the within-wave edit checks: depending on the wave and income type, 1-2% of total income sources were reported in response to the within-wave check, while 5-12% were reported in response to the RDI check. Their relative effects may be quite different if compared individually or in different order. Since the edit checks do not require feeding forward information from previous interviews, they can be used in cross-sectional surveys and are cheaper to implement than RDI. Their use is however restricted to income sources for which there are questions earlier in the questionnaire that are good predictors of eligibility.

The long-term effects of RDI have not been assessed. The ability of RDI to reduce underreporting is limited by the fact that the respondent can only be reminded of income sources reported in the past. Since RDI reduces underreporting in a given wave, this means that in the following wave a larger proportion of recipients can be reminded of income sources they have received in the past, increasing the effectiveness of RDI. As a result, over time RDI may decrease measurement error more than it does in a single wave. The effects of RDI across waves have not been assessed, as most previous studies have focused on the wave when dependent interviewing was first introduced. The effectiveness of the reminders depends on the quality of the first report and there is concern that dependent interviewing may lead measurement error to be fed forward into future waves. In the case of reporting of income sources, where the questions are about whether or not a source was received (yes/no), this risk is somewhat reduced. If the previous wave report was wrong and the respondent had underreported receipt, the RDI follow-up is not triggered in the following wave. That is, RDI simply has no effect. If the previous wave report is wrong because the respondent reported a source they had not actually received, the RDI follow-up would be triggered, leading to a risk that the respondent may continue to overreport receipt. The results of our validation study (see Table 6) however suggest that overreporting hardly occurs. Therefore we would conclude that underreporting in the previous wave reduces the effectiveness of RDI because the RDI check is not asked when it should be, but that overreporting in the previous wave does not impact the effectiveness of RDI.

The extent of measurement error in independent survey questions is presumably affected by the question format. The shortcut method of using showcards instead of separate yes/no questions about the receipt of all potential income sources presumably leads to more underreporting. On the other hand, the shorter interview time reduces respondent burden, which could lead to less measurement error using the showcards. This trade-off between cost savings in terms of questionnaire time and measurement error has not been assessed to our knowledge.

Finally, we have not touched on the question through which mechanisms RDI and edit checks work, that is, which types of sources are most likely to be misreported, by which types of respondents, and how the edit checks work for these different groups (see Lynn et al. 2012; Pascale et al. 2009). We have also not touched on the question how these methods could further be improved. Improvements could focus further on the reduction of underreporting, but also on capturing new receipt. This could be done by extending the use of within-wave edit checks by incorporating more factual questions into earlier sections of the questionnaire that predict eligibility for income receipt. Measurement error in household income was reduced by our study design, but there is room for further reductions in error with potentially greater impact on substantive conclusions.

5. References

AAPOR (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Reports for Surveys, (7th edition). Lenexa, KS: American Association for Public Opinion Research. Available at: http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156 (accessed October 30, 2013).

Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement Error in Survey Data. In Handbook of Econometrics Vol. 5, J.J. Heckman and E. Leamer (eds). Amsterdam: Elsevier, 3705–3843.

Brewer, M., Muriel, A., Philips, D., and Sibieta, L. (2009). Poverty and Inequality in the UK: 2009. The institute for fiscal studies, London. Available at: http://www.ifs.org.uk/comms/c109.pdf (accessed October 30, 2013).

Böheim, R. and Jenkins, S.P. (2006). A Comparison of Current and Annual Measures of Income in the British Household Panel Study. Journal of Official Statistics, 22, 733–758.

Hale, A. and Michaud, S. (1995). Dependent Interviewing: Impact on Recall and on Labour Market Transitions. SLID Research Paper Series No. 95-06. Ottawa: Statistics Canada.

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A., and Sala, E. (2006). Patterns of Consent: Evidence from a General Household Survey. Journal of the Royal Statistical Society Series A, 169, 701–722. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2006.00417.x

Jenkins, S.P., Lynn, P., Jäckle, A., and Sala, E. (2008). The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence from Britain. International Journal of Social Research Methodology, 11, 29–43. DOI: http://www.dx.doi.org/10.1080/13645570701401602

Jäckle, A. (2008). Measurement Error and Data Collection Methods: Effects on Estimates from Event History Data. ISER Working Paper 2008-13. Colchester: University of Essex. Available at: http://www.iser.essex.ac.uk/publications/working-papers/iser/2008-13.pdf (accessed October 30, 2013).

Jäckle, A. (2009). Dependent Interviewing: A Framework and Application to Current Research. In Methodology of Longitudinal Surveys, P. Lynn (ed.). Chichester: Wiley, 92–112.

Jäckle, H., Laurie, H., and Uhrig, S.C.N. (2007). The Introduction of Dependent Interviewing on the British Household Panel Survey. ISER Working Paper 2007-7. Colchester: University of Essex. Available at: http://www.iser.essex.ac.uk/publications/working-papers/iser/2007-07.pdf (accessed October 30, 2013).

Jäckle, A., Sala, E., Jenkins, S.P., and Lynn, P. (2004). Validation of Survey Data and Employment: the ISMIE Experience. ISER Working Paper 2004-14. Colcheste: University of Essex. Available at: http://www.iser.essex.ac.uk/publications/working-papers/iser/2004-14 (accessed October 30, 2013).

Lynn, P., Jäckle, A., Jenkins, S.P., and Sala, E. (2006). The Effects of Dependent Interviewing on Responses to Questions on Income Sources. Journal of Official Statistics, 22, 357–384.

Lynn, P., Jäckle, A., Jenkins, S.P., Sala, E. (2012). The Impact of Interviewing Method on Measurement Error in Panel Survey Measures of Benefit Receipt: Evidence from a Validation Study. Journal of the Royal Statistical Society, Series A, 175, 289-308. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2011.00717.x

Lugtig, P. and Jäckle, A. (2011). Can I just check. . . ? Effects of edit check questions on measurement error and survey estimates, ISER Working Paper 2011-23. Colchester: University of Essex. Available at: http://www.iser.essex.ac.uk/publications/working-papers/iser/2011-23 (accessed October 30, 2013).

Mathiowetz, N.A. and McGonagle, K.A. (2000). An Assessment of the Current State of Dependent Interviewing in Household Surveys. Journal of Official Statistics, 16, 401–418.

Moore, J., Bates, N., Pascale, J., and Okon, A. (2009). Tackling Seam Bias through Questionnaire Design. In Methodology of Longitudinal Surveys, P. Lynn (ed.). Chichester: Wiley, 72–92.

Pascale, J., Roemer, M.I., and Resnick, M. (2009). Medicaid Underreporting in the CPS. Results from a Record Check Study. Public Opinion Quarterly, 73, 497–520. DOI: http://www.dx.doi.org/10.1093/poq/nfp028

Pennell, S.G. (1993). Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income and Program Participation. Washington, DC: US Bureau of the Census. Available at: http://www.census.gov/sipp/workpapr/wp9314.pdf (accessed October 30, 2013).

Sala, E., Uhrig, S.C.N., and Lynn, P. (2009). It is Time Computers Do Clever Things! The Impact of Dependent Interviewing on Interviewer Burden. Field Methods, 23, 3–23. DOI: http://www.dx.doi.org/10.1177/1525822X10384087

Taylor, M.F., Brice, J., Buck, N., and Prentice-Lane, E. (2009). British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: University of Essex. Available at: http://www.iser.essex.ac.uk/bhps/documentation/pdf_versions/volumes/bhpsvola.pdf (accessed October 30, 2013).

Received December 2011
Revised May 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 63–90, http://dx.doi.org/10.2478/jos-2014-0004

Evaluation of Generalized Variance Functions in the Analysis of Complex Survey Data

MoonJung Cho[1], *John L. Eltinge*[1], *Julie Gershunskaya*[1], *and Larry Huff*[1]

Two sets of diagnostics are presented to evaluate the properties of generalized variance functions (GVFs) for a given sample survey. The first set uses test statistics for the coefficients of multiple regression forms of GVF models. The second set uses smoothed estimators of the mean squared error (MSE) of GVF-based variance estimators. The smooth version of the MSE estimator can provide a useful measure of the performance of a GVF estimator, relative to the variance of a standard design-based variance estimator. Some of the proposed methods are applied to sample data from the Current Employment Statistics survey.

Key words: Complex sample design; degrees of freedom; design-based inference; model-based inference; quarterly census of employment and wages; superpopulation model; U.S. current employment statistics (CES) survey; variance estimator stability.

1. Introduction

In the analysis of sample survey data, statisticians generally prefer to use variance estimation and inference methods that account for the complex design used in the selection of sample units. However, in some cases (especially those involving relatively small domains or other specialized subpopulations), standard design-based variance estimators may be unstable. For such cases, some analysts prefer to use "generalized variance functions" estimators, in which one seeks to approximate the true design or design-model variance as a function of known predictors X.

For some background on generalized variance functions for survey data, see Johnson and King (1987), Valliant (1987) and the references cited therein. (Some of this literature discusses other reasons for use of GVFs, for example, simplicity of use for secondary data analysts. The remainder of this article will not consider these other reasons in further detail.) Much of the GVF literature has focused on the variances of point estimators of population proportions or population totals related to a binary outcome variable (see, e.g., Bureau of Labor Statistics 2006, pp. 189–193). The current article, however, considers the more complex setting in which the point estimator of interest depends primarily on survey variables that are not binary. For example, the Current Employment Statistics survey

[1] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington, DC 20212, U.S.A. Emails: Cho.Moon@bls.gov, Eltinge.John@bls.gov, Gershunskaya.Julie@bls.gov, and Huff.Larry@bls.gov

Acknowledgments: The authors thank Pat Getz and Ken Robertson for many helpful discussions of the CES and an associate editor and three referees for their insightful suggestions on earlier versions of this article. The views expressed in this article are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

application in Subsection 2.1 and Section 5 depends on unit-level employment count reports that may range from one to tens of thousands.

Following the introduction of an illustrative example and a development of notation and prospective models in Section 2, this article develops two sets of diagnostic tools for GVFs. First, Section 3 presents design-based estimators of the variance-covariance matrix of the coefficient estimators for a GVF. The covariance-matrix estimators in turn lead to construction of test statistics and confidence sets for the GVF coefficients under standard large-sample conditions. Second, Section 4 develops diagnostics for the mean squared error of a GVF as an estimator of the true design variance of a given point estimator. An initial development reviews the relative magnitudes of error terms associated, respectively, with pure sampling variability of the design-based variance estimators; the deterministic lack of fit in the proposed GVF model; and the random equation error associated with the GVF model. Subsection 4.4 characterizes the unbiased MSE estimators of the GVF-based variance estimators in terms of the direct variance estimators. Subsection 4.5 fits models of these MSE estimators; produces a smooth version of the MSE estimators; and presents some simple methods of evaluating the relative magnitudes of the sampling error and equation error terms. Section 5 applies the proposed diagnostics to data from the U.S. Current Employment Statistics survey. Section 6 presents a simulation study that evaluates the properties of GVF coefficient estimators and of the related predictors of the true design variance. Section 7 summarizes the main ideas of this article and outlines some possible extensions. In addition, Table 1 provides a summary of the notation used in this article.

2. Illustrative Example, Background, Notation, and GVF Models

2.1. Illustrative Example: Subpopulation Total Estimators for the U.S. Current Employment Statistics Survey

The CES survey collects data monthly on employment, hours, and earnings from nonfarm establishments. Employment is the total number of persons employed full or part time in a nonfarm establishment during a specified month. One important feature of the CES survey is that complete universe employment counts of the previous year become available from the Unemployment Insurance (UI) tax records on a lagged basis (Butani et al. 1997). U.S. Bureau of Labor Statistics (2011, Ch. 2) describes the design features relevant to the analysis of the historical data considered in this article.

The CES sample design uses stratified sampling of UI accounts. UI account is a cluster that may contain a single or multiple establishment(s). An establishment is defined to be an economic unit, generally located at a single place, which is engaged predominantly in one type of economic activity. All establishments within a sampled UI account are included in the sample. When establishments are rotated into the sample, they are retained for two years or more. The strata are defined by state, industry, and the size class of UIs. The sample units in areas within each stratum are sorted in a way ensuring that the number of sampled units in each area is proportional to the area's size (i.e., proportional to the number of UIs in the frame for a given stratum).

Table 1. Description of notation

Notation	Description
b	index for elements of the coefficient vector γ
B	dimensionality of γ
C	dimensionality of ω
D	set of all j distinct domains
d_{jt}	degrees of freedom associated with the design-based distribution of V_{pjt}
$d*$	degrees of freedom associated with the superpopulation distribution of $(V^*_{pjt})^{-1}V_{pjt}$
d_w	degrees of freedom in the Wishart distribution for $\hat{V}\hat{w}(\hat{\gamma})$
h_f	smooth version of $E\{(V^*_{pjt} - V_{pjt})^2 \mid X_{jt}\}$
i	industry
j	domain
n_{jt}	number of responding sample UI accounts in domain j at time t
p	sample design
q_{jt}	equation error
r_{jt}	residuals with expectation $E(q^2_{jt}\mid X_{jt})$
$\hat{R}$	growth ratio estimate
$SE1$	square root of $(2\hat{V}^2_{pjt})/(d_{jt}+2)$
$SE2$	square root of $(2V^{*2}_{pjt})/d$
t	months from benchmark month
V_{pjt}	design variance of $\hat{\theta}_{jt}$
$\hat{V}_{pjt}$	variance estimator based on the design
V^*_{pjt}	variance estimator based on the model
X	vector of predictor variables for GVF model
y	unknown true employment total
Z	vector of predictor variables for the residual Models (21) through (24)
γ	variance function parameters in Model (f)
ε_{jt}	sampling error $\hat{V}_{pjt} - V_{pjt}$
η_{jt}	error term in Model (22)
θ_{jt}	finite population quantity
$\theta_{\xi jt}$	superpopulation analogue of θ_{jt}
$\hat{\theta}_{jt}$	point estimator of θ_{jt}
ξ	superpopulation index
$\hat{\sigma}^2_e$	residual mean squared error terms
ω	variance function parameters

For this article, the survey variable of main interest is y_{jtk}, defined to equal the total employment reported by establishment k within domain j for reference month t. The universe data, known as Quarterly Census of Employment and Wages (QCEW) data, are used annually to benchmark the CES sample estimates to these universe counts (Werking 1997). Specifically, let x_{j0} equal the known QCEW employment total within domain j for the benchmark month 0. In addition, let y_{jt} equal the unknown true employment total for domain j in month t. CES uses a "weighted link relative estimator" of y_{jt}, computed as

the product,

$$\hat{y}_{jt} = x_{j0}\hat{R}_{jt},$$

where $\hat{R}_{jt}$ is an estimator of the relative employment growth that took place from benchmark month 0 to the current month t. Specifically,

$$\hat{R}_{jt} = \prod_{\tau=1}^{t} \hat{R}^{*}_{j\tau},$$

where $\hat{R}^{*}_{j\tau} = \left(\sum_{k\in s_{j\tau}} w_k y_{jk,\tau-1}\right)^{-1} \sum_{k\in s_{j\tau}} w_k y_{jk\tau}$, $s_{j\tau}$ is the matched sample of establishments in domain j that report positive employment in both months $\tau - 1$ and τ, and w_k is the sampling weight of establishment k. Note especially that $\hat{R}_{jt}$ equals the product of t separate estimators of one-month change. Consequently, under regularity conditions, one may anticipate that $\hat{R}_{jt}$ and $\hat{y}_{jt}$ may have design variances that are increasing functions of t. For more detailed information on the weighted link relative estimator, see BLS Handbook of Methods (2011) and Gershunskaya and Lahiri (2005). For data used in this article, the benchmark month ($t = 0$) is March 1999 and our sample data will lead to employment estimates for each month from January through December 2000 ($t = 10$ to $t = 21$).

The primary CES design goal is to satisfy the precision requirements specified for the national estimates. However, there is strong substantive interest in finer domains which are defined by geographic characteristics and industrial classifications. For example, the data analyses in Section 5 focus on estimates of total employment for 430 domains defined by the intersection of metropolitan statistical area (MSA) with industry, for example, durable goods manufacturing in the St. Louis MSA or wholesale trade in the Charleston MSA. Within these domains, effective sample sizes become so small that the standard design-based estimators are not precise enough to satisfy the needs of prospective data users (Eltinge et al. 2001). It is necessary to have stable estimators of $V(\hat{y}_{jt})$ for the finer domains. Consequently, we considered the use of GVF methods to produce domain-level variance estimators that would be more stable than direct design-based variance estimators.

2.2. *Background and Notation*

Let θ_{jt} be a finite population mean or total for period t, and let $\theta_{\xi jt}$ be a superpopulation analogue of θ_{jt} where j is the domain index. For example, in the CES survey, domains are the combinations of industries and areas, and are generally studied for a sequence of months $t = 1, \ldots, T$. In addition, let $\hat{\theta}_{jt}$ be a point estimator of θ_{jt}; and define $V_{pjt} = V_p(\hat{\theta}_{jt})$ to be the design variance of $\hat{\theta}_{jt}$. Throughout this article, the subscript "p" denotes an expectation or variance evaluated with respect to the sample design. The GVF models the variance of a survey estimator, V_{pjt}, as a function of the parameter θ_{jt} and possibly other variables (Wolter 2007, sec. 7.2). A common specification is

$$V_{pjt} = f(X_{jt,}\gamma) + q_{jt}, \tag{1}$$

where X_{jt} is a vector of predictor variables potentially relevant to estimators of V_{pjt}, q_{jt} is a random univariate "equation error" with the mean 0, and γ is a vector of B-dimensional variance function parameters which we need to estimate. Note especially that q_{jt}

represents the deviation of the true design variance V_{pjt} from its modeled value $f(X_{jt}, \gamma)$. One generally would view the error term q_{jt} as arising from the superpopulation model that generated our finite population.

In some GVF applications, one may consider functions $f(\cdot)$ that depend on the domain-specific parameter θ_{jt} and may also consider cases for which some predictors X_{jt} are unknown and replaced by estimated terms, say $\hat{X}_{jt}$. However, these cases did not arise in the CES application considered here, so this article will limit its attention to forms of the Model (1) with known predictors X_{jt}.

In general, it is not possible to observe the true design variance V_{pjt}. Instead it is possible to compute an estimator $\hat{V}_{pjt} = \hat{V}_p(\hat{\theta}_{jt})$ based on, for example linearization or replication-based methods. Consequently, Model (1) must be supplemented with the decomposition

$$\hat{V}_{pjt} = V_{pjt} + \epsilon_{jt}, \tag{2}$$

where ϵ_{jt} is a random term that reflects sampling error in the estimator $\hat{V}_{pjt}$. Under the assumption that $\hat{V}_{pjt}$ is design unbiased for V_{pjt}, the error term ϵ_{jt} has design expectation equal to zero. The distinction between the equation error in Model (1) and the sampling error in Model (2) has been considered in other settings for analysis of experiments with replicates (e.g., Draper and Smith 1998, p. 47) and measurement error models (e.g., Fuller 1987).

Our CES applications will use a special form of Model (1) on the logarithmic scale,

$$\ln(V_{pjt}) = X_{jt}\gamma + q^*_{jt}, \tag{3}$$

where q^*_{jt} is a general error term with mean equal to zero; Appendix C provides some related details. A relatively simple form of Model (3) that incorporates factors related to domain size (x_{j0}), number of respondents (n_{jt}) and distance from benchmark month 0 to the reference period (t) is:

$$\ln(V_{pjt}) = \gamma_0 + \gamma_1 \ln(x_{j0}) + \gamma_2 \ln(n_{jt}) + \gamma_3 \ln(t) + q^*_{jt}. \tag{$f1$}$$

To estimate the parameters of Models (2) and (3), let D be the set of all J distinct domains (area-industry combinations) and for each $j \in D$, let D_{jt} be the set of responding sample establishments in domain j for month t. In addition, let $\mathbf{Y}_j$ be a $T \times 1$ vector with t-th element $\ln(\hat{V}_{jt})$ and define the $(J \cdot T) \times 1$ vector $\mathbf{Y} = (\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_J)'$. Similarly, let $\mathbf{X}_j$ be a $T \times B$ matrix with t-th row $\mathbf{X}_j(t, :)$ equal to the predictors used for the specified GVF model. Also, define the $(J \cdot T) \times B$ matrix $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_J)'$ and $B \times 1$ vector $\gamma = [\gamma_1, \ldots, \gamma_B]'$. For example, under the Model ($f1$), $\mathbf{X}_j(t, :) = [1, \ln(x_{j0}), \ln(n_{jt}), \ln(t)]$ and $\gamma = [\gamma_0, \gamma_1, \gamma_2, \gamma_3]'$. Then one may compute the ordinary least squares estimator of the coefficient vector in (3) as

$$\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4}$$

2.3. GVF Models

We used the logarithms of direct variance estimators $\hat{V}_{pjt}$ from the survey as the dependent variables in GVF models. The CES data we considered were from reference year 2000, and direct estimators, $\hat{V}_{pjt}$ of V_{pjt}, were computed from Fay's variant of the balanced

half-sample replication method with adjustment term $K = 0.5$. For general background on balanced half-sample replication and Fay's method, see Wolter (2007, Ch. 3) and Judkins (1990). For sampling within a given industry, the CES uses eight size classes. For variance estimation, the CES combines the three largest size classes (6, 7 and 8). So there are six size-based variance strata within each area-industry domain.

We assume that $\hat{V}_{pjt}$ is a design-unbiased estimator for V_{pjt}, i.e., $E_p(\hat{V}_{pjt}) = V_{pjt}$. Let n_{jt} be the number of responding sample UI accounts within the domain j and month t. In this article, we consider only domains with at least twelve reporting UI accounts. There are 430 area-industry combinations in our CES data. Each area-industry combination has data from January to December of the year 2000. For the current analysis, we considered data from the six industries described in Table 2. For areas with a substantial amount of mining activity, CES produces separate employment estimates for the mining and construction industries respectively. For other areas, CES produces a single employment estimate for the combined mining and construction industries. For the 430 domains considered here, the mean number of reporting sample UI accounts was 475. For the CES application, this article will consider three special cases of Model (1) on a logarithmic scale. First, note that Model ($f1$) from Subsection 2.2 constrains both intercepts and slopes to be constant across industries and areas. A generalization that allows the intercepts to vary across industries is:

$$\ln(V_{jt}) = \gamma_{0i(j)} + \gamma_1 \ln(x_{j0}) + \gamma_2 \ln(n_{jt}) + \gamma_3 \ln(t) + q^*_{jt}, \qquad (f2)$$

where $i(j)$ represents the industry i that is represented in a specific domain j. A further generalization that allows all coefficients to vary across industries is:

$$\ln(V_{jt}) = \gamma_{0i(j)} + \gamma_{1i(j)} \ln(x_{j0}) + \gamma_{2i(j)} \ln(n_{jt}) + \gamma_{3i(j)} \ln(t) + q^*_{jt}. \qquad (f3)$$

Thus, Model ($f3$) allows interaction between the industry classification and the predictors x_{j0}, n_{jt} and t. Note that in the notation of the general expression (1), Models ($f1$) through ($f3$) involve only predictors X determined by the respondent count n_{jt}, the time lag t and the terms x_{j0}. In contrast with GVFs used for binary outcome variables (e.g., Johnson and King 1987), Models ($f1$) through ($f3$) do not use the population parameters θ_{jt} as scale factors. Instead, our models use the known benchmark total x_{j0} as the scale-factor predictor. Also, for each industry considered in Model ($f3$), we used data from twelve months and from two to 131 areas, as specified in Table 2. In addition, Wolter (2007, Sec. 7.3) and others have noted the importance of fitting GVF models for groups of

Table 2. Number of metropolitan areas (MSAs) and UIs in each industry

Industry description	MSAs	Sample UIs
1 Mining	2	549
2 Mining and construction	36	22,359
3 Construction	61	54,552
4 Durable manufacturing	131	76,150
5 Nondurable manufacturing	100	50,717
6 Wholesale trade	100	58,424

statistics $\hat{\theta}_{jt}$ for which a "common model" will hold. Model ($f1$) uses a common model for all domains (j), while Model ($f3$) has distinct coefficient vectors γ for each industry i. In other words, Model ($f1$) uses a single large "group" while Model ($f3$) allows each industry to be a separate group.

3. Estimation and Inference for Coefficients in a GVF Model

3.1. Point Estimation Methods

For each of Models ($f1$) through ($f3$), we computed estimators $\hat{\gamma}$ of the coefficients γ through ordinary least squares (OLS) regression of $\ln(\hat{V}_{jt})$ on the corresponding vector of predictors. In principle, one could consider alternative coefficient estimators based on weighted least squares or generalized least squares approaches. However, the efficiency gains from these alternative approaches, if any, would depend on the covariance structure of the error terms; details will not be considered in the current article. See Valliant (1987) for a discussion of conditions under which weighted least squares estimation may be preferred to ordinary least squares estimation for GVFs.

3.2. Design-Based Variance Estimation for GVF Coefficients

We obtain an estimator $\hat{V}_p(\hat{\gamma})$ of the variance of the approximate distribution of $\hat{\gamma}$ from an extension of standard estimating equation approaches for complex-survey estimators (Binder 1983). Then the estimator $\hat{\gamma}$ in Expression (4) can be rewritten as the solution of the estimating equation,

$$\begin{aligned} 0 &= \hat{\mathbf{w}}(\gamma) \\ &= \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\gamma \\ &= \sum_{j\in\mathcal{D}} \hat{\mathbf{w}}_{j\cdot}(\gamma), \end{aligned}$$

where $\hat{\mathbf{w}}_{j\cdot}(\gamma) = \mathbf{X}_j'(\mathbf{Y}_j - \mathbf{X}_j\gamma)$. In addition, let $\hat{\mathbf{w}}_{jb}(\gamma)$ be the b-th element of $\hat{\mathbf{w}}_{j\cdot}(\gamma)$ and let $\hat{\mathbf{w}}_{\cdot b}(\gamma)$ be the b-th element of $\hat{\mathbf{w}}(\gamma)$. The Taylor expansion of $\hat{\mathbf{w}}(\gamma)$ at $\gamma = \gamma^*$, where γ^* is the population parameter value, leads to:

$$\begin{aligned} 0 &= \hat{\mathbf{w}}(\hat{\gamma}) \\ &= \hat{\mathbf{w}}(\gamma^*) + \hat{\mathbf{w}}^{(1)}(\gamma^*)(\hat{\gamma} - \gamma^*) + R, \end{aligned}$$

where $\hat{\mathbf{w}}^{(1)}(\gamma^*) = \frac{\partial \hat{\mathbf{w}}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma^*}$ and R is a $B \times 1$ vector with b-th element equal to $2^{-1}(\hat{\gamma} - \gamma^*)'\left(\frac{\partial^2 \hat{\mathbf{w}}_{\cdot b}(\gamma)}{\partial\gamma\partial\gamma'}\Big|_{\gamma=\gamma^{**}}\right)(\gamma - \gamma^*)$ for some γ^{**} with $|\gamma^{**} - \gamma^*| < |\hat{\gamma} - \gamma^*|$. Thus,

$$\hat{\mathbf{w}}(\gamma^*) = -\hat{\mathbf{w}}^{(1)}(\gamma^*)(\hat{\gamma} - \gamma^*) - R. \tag{5}$$

Under regularity conditions, the second term on the right-hand side of Expression (5) is of a smaller order of magnitude than the first term. Consequently, an estimator of the

variance-covariance matrix of the approximate distribution of $\hat{\gamma}$ is

$$\hat{V}(\hat{\gamma}) = \left\{\hat{\mathbf{w}}^{(1)}(\hat{\gamma})\right\}^{-1} \hat{V}\{\hat{\mathbf{w}}(\hat{\gamma})\}\left[\left\{\hat{\mathbf{w}}^{(1)}(\hat{\gamma})\right\}\right]^{-1}, \tag{6}$$

where $\hat{\mathbf{w}}^{(1)}(\hat{\gamma}) = \frac{\partial \hat{\mathbf{w}}(\gamma)}{\partial \gamma}\Big|_{\gamma=\hat{\gamma}}$ and $\hat{V}\{\hat{\mathbf{w}}(\hat{\gamma})\}$ is an estimator of the variance of $\hat{\mathbf{w}}(\hat{\gamma})$, evaluated at the point $\gamma = \hat{\gamma}$.

3.3. *Application to the Current Employment Statistics Program*

Let T be the total number of months covered by the data; for the CES design, $T = 12$. Then

$$\hat{\mathbf{w}}^{(1)}(\hat{\gamma}) = -\sum_{j\in\mathcal{D}}\sum_{t=1}^{T} \boldsymbol{X}'_{jt}\boldsymbol{X}_{jt}.$$

In addition, under the CES design, selection of sample units is essentially independent across domains. However, due to the CES design and estimation methods, estimators within a domain may be strongly correlated across consecutive months. Consequently, an estimator for the middle term in Expression (6) is

$$\begin{aligned}\hat{V}\{\hat{\mathbf{w}}(\hat{\gamma})\} &= \hat{V}\left(\sum_{j\in\mathcal{D}} \hat{\mathbf{w}}_j(\hat{\gamma})\right) \\ &= J^2\hat{V}\left(J^{-1}\sum_{j\in\mathcal{D}} \hat{\mathbf{w}}_j(\hat{\gamma})\right) \\ &= (J-1)^{-1}J\sum_{j\in\mathcal{D}}\left\{\hat{\mathbf{w}}_j(\hat{\gamma}) - \hat{\bar{\mathbf{w}}}(\hat{\gamma})\right\}\left\{\hat{\mathbf{w}}_j(\hat{\gamma}) - \hat{\bar{\mathbf{w}}}(\hat{\gamma})\right\}',\end{aligned} \tag{7}$$

where $\hat{\bar{\mathbf{w}}} = J^{-1}\sum_{j\in D}\hat{\mathbf{w}}_j(\hat{\gamma})$. Note that the final equality in Expression (7) uses the independence across domains j and accounts for correlation across periods t. Under regularity conditions (e.g., Korn and Graubard 1990) $d_w\hat{V}(\hat{\gamma})$ is distributed approximately as a Wishart random matrix on d_w degrees of freedom and matrix parameter $V(\hat{\gamma})$.

4. Comparison of the Direct and GVF Methods in Prediction of the True Variance

4.1. *Decomposition of Differences of* $\hat{V}_{pjt} - V^*_{pjt}$

Once we have selected and estimated a specific GVF Model (f), it is useful to evaluate the properties of the resulting predictors of V_{pjt}. Suppose that a model-fitting method (e.g., ordinary least squares, perhaps on a transformed scale; or nonlinear least squares) leads to the coefficient point estimator $\hat{\gamma}$, and define the resulting variance terms,

$$V^*_{pjt} \stackrel{def}{=} f(X_{jt}, \hat{\gamma}). \tag{8}$$

Appendix C presents two options for specific ways in which to incorporate parameter estimators into the adjusted predictors V^*_{pjt}. The data analysis for this article will use a fairly conservative predictor V^*_{pjt}. Note that V^*_{pjt} is based on the general model (1) given

on the original variance scale. Under the variance function model (1), error model (2) and the definition of V^*_{pjt} in Expression (8), $\hat{V}_{pjt} - V_{pjt} = \epsilon_{jt}$, and $V^*_{pjt} - V_{pjt} = f(X_{jt}, \hat{\gamma}) - \{f(X_{jt}, \gamma) + q_{jt} + E(q_{jt}) - E(q_{jt})\}$. Consequently, we may decompose the difference $\hat{V}_{pjt} - V^*_{pjt}$ as

$$\begin{aligned} \hat{V}_{pjt} - V^*_{pjt} &= (\hat{V}_{pjt} - V_{pjt}) - (V^*_{pjt} - V_{pjt}) \\ &= \epsilon_{jt} + \{q_{jt} - E(q_{jt})\} + E(q_{jt}) - \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}. \end{aligned} \tag{9}$$

In Equation (9), ϵ_{jt} is a pure estimation error in the original $\hat{V}_{pjt}$ estimates with $E(\epsilon_{jt}) = 0$; $\{q_{jt} - E(q_{jt})\}$ is a random equation error; and $E(q_{jt})$ represents the deterministic lack-of-fit in our model attributable, for example, to omitted regressors or a misspecified functional form. The last term in Equation (9), $\{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}$, is a parameter estimation error attributable to the errors $\hat{\gamma} - \gamma$.

Exploratory analysis of the adequacy of our estimated values, V^*_{pjt}, may focus on the magnitude of the prediction errors $\left(V^*_{pjt} - V_{pjt}\right)$, relative to the errors $\left(\hat{V}_{pjt} - V_{pjt}\right)$, in the original estimators $\hat{V}_{pjt}$. If $E\left(V^*_{pjt} - V_{pjt}\right)^2$ is smaller than the variance of $\hat{V}_{pjt}$, then we would prefer V^*_{pjt}. In addition,

$$\delta(X_{jt}, \gamma) \stackrel{def}{=\!=} E[\{f(X_{jt}, \hat{\gamma}) - V_{pjt}\}^2 | X_{jt}, \gamma]$$

may vary across values of X_{jt} with $\delta(X_{jt}, \gamma) << V_p(\hat{V}_{pjt} - V_{pjt})$ only in some cases. In this case, we might prefer V^*_{pjt} for some, but not all values of X_{jt}.

4.2. *Properties of the Direct Estimator* $\hat{V}_{pjt}$

We evaluate error sizes in terms of conditional expected squared error. In keeping with standard evaluation of design-based variance estimators, assume that for positive d_{jt},

$$E_p(\hat{V}_{pjt} | V_{pjt}) = V_{pjt}, \quad V_p(\hat{V}_{pjt} | V_{pjt}) = \frac{2V^2_{pjt}}{d_{jt}}. \tag{10}$$

The moment properties (10) would hold if $V^{-1}_{pjt} d_{jt} \hat{V}_{pjt}$ followed a $\chi^2(d_{jt})$ distribution. However, the current article will assume that $\hat{V}_{pjt}$ follows a lognormal distribution that in general would allow somewhat greater modeling flexibility; see Appendix B for related comments. Note that

$$\begin{aligned} E_p(\hat{V}^2_{pjt} | V_{pjt}) &= \{E_p(\hat{V}_{pjt}) | V_{pjt}\}^2 + V_p(\hat{V}_{pjt} | V_{pjt}) \\ &= V^2_{pjt} + \frac{2V^2_{pjt}}{d_{jt}} \\ &= d^{-1}_{jt}(d_{jt} + 2)V^2_{pjt}. \end{aligned} \tag{11}$$

Consequently from (11), an unbiased estimator of $V_p(\hat{V}_{pjt} | V_{pjt})$ is:

$$\hat{V}_p(\hat{V}_{pjt} | V_{pjt}) = (d_{jt} + 2)^{-1} 2\hat{V}^2_{pjt}. \tag{12}$$

Six employment size classes were used for stratification for our CES survey example, so the data analysis in Section 5 will use $d_{jt} = 6$. In addition, sample sizes within

employment class generally were large enough for each t that stratum-level sample means were considered to follow an approximate normal distribution.

4.3. Properties of the GVF Estimator V^*_{pjt}

Now consider the properties of V^*_{pjt}, and the conditions under which V^*_{pjt} may have a smaller mean squared error than $\hat{V}_{pjt}$. In the general case,

$$V_{pjt} - V^*_{pjt} = q_{jt} - \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}. \tag{13}$$

To simplify the discussion, assume that the product $(J \cdot T)$ is increasing without bound. This would occur with, for example, increases in the number of geographical areas or the number of time periods. For example, the CES application uses data from 430 area-industry combinations and 12 time periods, so the product $J \cdot T$ is relatively large. Then, under mild regularity conditions on the function $f(\cdot)$,

$$E[\{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}^2 | X_{jt}] = \mathbf{O}_p\{(J \cdot T)^{-1}\}, \tag{14}$$

while the domain-specific term $E\left(q^2_{jt}|X_{jt}\right)$ does not necessarily decrease as the product $(J \cdot T)$ increases. For example, result (14) generally holds for each of Models $(f1)$-$(f3)$ because these models do not include terms θ_{jt}; include only known predictors X_{jt}; and involve errors $\hat{\gamma} - \gamma$ that are $\mathbf{O}_p\{(J \cdot T)^{-1/2}\}$. Under result (14) and additional technical conditions,

$$E\{(V^*_{pjt} - V_{pjt})^2 | X_{jt}\} = E(q^2_{jt}) + \mathbf{O}_p\{(J \cdot T)^{-1}\} \tag{15}$$

and the leading term $E(q^2_{jt})$ will generally be of larger magnitude than the $\mathbf{O}_p\{(J \cdot T)^{-1}\}$ term associated with the error $f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)$. Consequently, our task of evaluation of the approximate mean squared error of V^*_{pjt} simplifies to an evaluation of the expected square of q_{jt}.

4.4. Diagnostics for Comparison of $\hat{V}_{pjt}$ And V^*_{pjt}

We do not observe q_{jt} directly, but we can estimate its expected square through the following steps. First, note from Expression (9) that

$$\hat{V}_{pjt} - V^*_{pjt} = \epsilon_{jt} + q_{jt} - \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}$$

and so

$$\begin{aligned}\left(\hat{V}_{pjt} - V^*_{pjt}\right)^2 = {} & \epsilon^2_{jt} + q^2_{jt} \\ & + \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}^2 \\ & + 2q_{jt}\left\{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\right\} \\ & + 2\epsilon_{jt}\left[q_{jt} - \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}\right].\end{aligned} \tag{16}$$

Under condition (14), the conditional expectation $E\left(\{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}^2 | X_{jt}\right)$ is small relative to $E(q^2_{jt}|X_{jt})$. Under additional mild conditions, the conditional expectations $E\left[2q_{jt}\left\{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\right\}|X_{jt}\right]$ and $E\left(2\epsilon_{jt}\left[q_{jt} - \{f(X_{jt}, \hat{\gamma}) - f(X_{jt}, \gamma)\}\right]|X_{jt}\right)$ are small

relative to $E\left(q_{jt}^2|X_{jt}\right)$, so

$$E\left\{\left(\hat{V}_{pjt} - V_{pjt}^*\right)^2|X_{jt}\right\} \doteq V_p(\hat{V}_{pjt}|V_{pjt}) + E\left(q_{jt}^2|X_{jt}\right). \tag{17}$$

Expressions (9) and (16) lead to two general conclusions regarding diagnostics for V_{pjt}^*. First, due to distinctions between $V(\epsilon_{jt})$ and $E(q_{jt}^2)$, care is required in the interpretation of standard regression diagnostics when applied to GVF models like the general model (1), or the specific models ($f1$) through ($f3$). For example, the customary regression mean squared error, $\hat{\sigma}_e^2$, is an estimator of the sum $V(\epsilon_{jt}) + E(q_{jt}^2)$. In addition, under regularity conditions, the customary squared coefficient of variation, R^2, satisfies the approximate relationship,

$$1 - R^2 \doteq \left\{V(\epsilon_{jt}) + E(q_{jt}^2) + (J-1)^{-1}\sum_{j=1}^{J}\left\{f(X_{jt}, \hat{\gamma}) - \bar{\hat{V}}\right\}^2\right\}^{-1} \left\{V(\epsilon_{jt}) + E(q_{jt}^2)\right\}, \tag{18}$$

where $\bar{\hat{V}} = J^{-1}\sum_{j=1}^{J}\hat{V}_{jt}$. Under an ideal fit for Model (1), $E(q_{jt}^2)$ would be approximately equal to zero, but $1 - R^2$ would not necessarily be close to zero, due to the presence of $V(\epsilon_{jt})$ in the numerator of Expression (18). Thus, relatively small values of R^2 by themselves do not necessarily indicate a poor fit for GVF Model (1). Similar comments apply to other regression goodness-of-fit diagnostics used for GVF models.

Second, to address these limitations, it is useful to consider estimators of $E(q_{jt}^2|X_{jt})$ and related diagnostics that adjust for the effects of $V(\epsilon_{jt})$. In particular, Expression (12) is an unbiased estimator of the first term on the right-hand side of Expression (17). Consequently, we may define a direct estimator of $E(q_{jt}^2|X_{jt})$ to be

$$r_{jt} \stackrel{def}{=\!=} \left(\hat{V}_{pjt} - V_{pjt}^*\right)^2 - (d_{jt}+2)^{-1}2\hat{V}_{pjt}^2. \tag{19}$$

Note that r_{jt} is a random variable with properties that depend on the distributions of both the equation error term q_{jt} and the sampling error term ϵ_{jt}. For example, if $E(q_{jt}^2|X_{jt}) = 0$, then the leading terms of a Taylor expansion of r_{jt} would have an expectation equal to zero. Similarly, if $E(q_{jt}^2|X_{jt})$ is not large relative to $E(\epsilon_{jt}^2|X_{jt})$, then there is a substantial probability that a given value of r_{jt} is less than zero. These results are similar to properties of unadjusted estimators of "between group" variance terms in standard variance component models. For example, for the data analysis detailed in Section 5, approximately 36% of the r_{jt} values were less than zero.

Consequently, in assessment of $E(q_{jt}^2|X_{jt})$, use of smoothed versions of r_{jt} would generally be preferred. For example, one could extend the standard variance-component literature on "restricted maximum likelihood" (REML) estimation (e.g., Patterson and Thompson 1971; Corbeil and Searle 1976; and Harville 1977). However, a detailed extension of REML methods to the current setting is beyond the scope of the current work. Instead, the next subsection presents a relatively simple regression approach to estimation of $E(q_{jt}^2|X_{jt})$.

4.5. Model Fitting for Conditional Expected Squared Equation Error

In general, one may consider a model

$$\begin{aligned} E(q_{jt}^2|X) &= Z_{jt}\omega + \eta_{jt} \\ &= \sum_{c=1}^{C} Z_{cjt}\omega_c + \eta_{jt} \end{aligned} \tag{20}$$

for the conditional expectation of q_{jt}^2, where $Z_{jt} = (Z_{1jt}, \ldots, Z_{Cjt})$ is a $1 \times C$ vector of predictors (generally functions of θ_{jt}, X_{jt} and γ); $\omega = (\omega_1, \ldots, \omega_C)'$ is a $C \times 1$ column of fixed regression coefficients; and η_{jt} is a random error term arising from the underlying superpopulation model. Let $\mathbf{r}_j$ be a $T \times 1$ vector with t-th element r_{jt} and define the $(J \cdot T) \times 1$ vector $\mathbf{r} = (\mathbf{r}_1', \mathbf{r}_2', \ldots, \mathbf{r}_J')'$. Similarly, let $\mathbf{Z}_j$ be a $T \times C$ matrix with t-th row $\mathbf{Z}_j(t,:)$ equal to the predictors used for the specified model. Also, define the $(J \cdot T) \times C$ matrix $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2', \ldots, \mathbf{Z}_J')'$. Define $\hat{\omega} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{r}$. See Appendix A for development of the variance estimators and inferential statistics for $\hat{\omega}$. Finally, define an estimator of $E(\mathbf{r}|\mathbf{Z})$ by

$$\hat{h}_f = \mathbf{Z}\hat{\omega}. \tag{21}$$

For example, in keeping with the general approach to error analysis in variance function models (e.g., Davidian et al. 1988), a quadratic function version of Model (20) is

$$V(q_{jt}|\theta_{jt}, X_{jt}, \gamma) = \omega_0 + \omega_1 f(X_{jt}, \gamma) + \omega_2 \left\{f(X_{jt}, \gamma)\right\}^2 + \eta_{jt}, \tag{22}$$

where $E(\eta_{jt}) = 0$. Under approximation (15) and Model (22), the relative variance of the prediction error $V_{pjt} - V_{pjt}^*$ is

$$\begin{aligned} &RelVar\left(V_{pjt} - V_{pjt}^*|X_{jt}\right) \\ &\doteq \left\{f(X_{jt}, \gamma)\right\}^{-2} V\left(V_{pjt} - V_{pjt}^*\right) \\ &\doteq \left\{f(X_{jt}, \gamma)\right\}^{-2}\omega_0 + \left\{f(X_{jt}, \gamma)\right\}^{-1}\omega_1 + \omega_2 + \left\{f(X_{jt}, \gamma)\right\}^{-2}\eta_{jt}. \end{aligned} \tag{23}$$

When condition (14) does not hold, one could consider an expansion of Model (22) to account for predictors of the additional components of $RelVar\left(V_{pjt} - V_{pjt}^*|X_{jt}\right)$. For a given function $f(X_{jt}, \gamma)$, we may consider a model to produce a smooth version, $h_f(X_{jt}, \omega)$, of the conditional expectation, $E\{(V_{pjt}^* - V_{pjt})^2|X_{jt}\}$, such that:

$$E\left\{\left(V_{pjt}^* - V_{pjt}\right)^2|X_{jt}\right\} = h_f(X_{jt}, \omega) + \eta_{jt}.$$

For example, Expression (22) leads to

$$r_{jt} = \omega_0 + \omega_1 V_{pjt}^* + \omega_2 V_{pjt}^{*2} + a_{jt}, \tag{24}$$

where we substitute the observed values V_{pjt}^* for the unknown quantities $f(X_{jt}, \gamma)$, and a_{jt} is a remainder term. In addition, it is of interest to consider the reduced form of Model (24)

in which $\omega_0 = 0 = \omega_1$:

$$\bar{V}^{*-2} r_{jt} = \bar{V}^{*-2} V_{pjt}^{*2} \omega_2 + \bar{V}^{*-2} a_{jt}, \tag{25}$$

where $\bar{V}^* = J^{-1} \sum_{j=1}^{J} V_{pjt}^*$. For example, under Model (24), $\mathbf{Z}_j(t, \cdot) = [1, V_{pjt}^*, V_{pjt}^{*2}]$ and $\omega = [\omega_0, \omega_1, \omega_2]'$ where j is the number of domains, and $C = 3$ is the number of coefficients in (24). Similarly, for Model (25), $\mathbf{Z}_j(t, \cdot) = \left[V_{pjt}^{*2}\right]$ and $C = 1$.

4.6. *A Degrees-of-Freedom Interpretation of Prediction Error Properties*

Application of the ideas in Appendix B indicate that under Model (22), the term

$$\left\{f(X_{jt}, \gamma)\right\}^{-1} d_{jt}^* V_{pjt} \tag{26}$$

has the same first and second moments as a $\chi^2_{d_{jt}^*}$ random variable where

$$\begin{aligned} d_{jt}^* &= 2\left\{RelVar(V_{pjt} - V_{pjt}^*)\right\}^{-1} \\ &\doteq \left[\left\{f(X_{jt}, \gamma)\right\}^{-2} \omega_0 + \left\{f(X_{jt}, \gamma)\right\}^{-1} \omega_1 + \omega_2 + \left\{f(X_{jt}, \gamma)\right\}^{-2} \eta_{jt}\right]^{-1} 2. \end{aligned} \tag{27}$$

In addition, under Model (24), results presented in Appendix B indicate that $2\left(V_{pjt}^{*-2} \hat{h}_f\right)^{-1}$ is an estimator of Expression (27) provided the error difference $V_{pjt}^{*-2} a_{jt} - V_{pjt}^{*-2} \eta_{jt}$ is small. Thus, the degrees-of-freedom attributable to the error term q_{jt} may in general depend on the function $f(X_{jt}, \gamma)$ and thus vary across domains.

However, under the reduced Model (25), if the remainder term a_{jt} is small, then

$$d_{jt}^* \doteq \omega_2^{-1} 2, \tag{28}$$

that is, the degrees-of-freedom term d_{jt}^* is approximately constant and can be estimated on the basis of the estimated coefficient ω_2 from the reduced Model (25).

5. Data Analysis

5.1. *Estimation for GVF Model Coefficients*

For the CES example introduced in Section 2, Tables 3 through 5 report coefficient estimates, standard errors and inferential statistics for Models $(f1)$ through $(f3)$ respectively. The reported standard errors equal the square root of the variance estimates

Table 3. Coefficient estimates and inferential statistics for Model (f1)

	Intercept γ_0	$\ln(x_{j0})$ γ_1	$\ln(n_{jt})$ γ_2	$\ln(t)$ γ_3	R^2	$\hat{\sigma}_e^2$
EST.	−1.43	1.16	0.22	1.17	0.52	1.31
s.e.	0.66	0.09	0.12	0.07		
t_γ	−2.17	12.77	1.78	16.72		
meff	5.45	9.87	10.63	1.02		

Table 4. Coefficient estimates and inferential statistics for Model (f2)

	Intercept						$\ln(x_{j0})$	$\ln(n_{jt})$	$\ln(t)$		
	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_1	γ_2	γ_3	R^2	$\hat{\sigma}_e^2$
EST.	−3.98	−3.28	−3.44	−4.85	−4.89	−4.26	1.70	−0.57	1.25	0.61	1.06
s.e.	1.08	0.65	0.66	0.72	0.73	0.71	0.11	0.15	0.07		
t_γ	−3.68	−5.03	−5.23	−6.76	−6.74	−5.99	16.06	−3.86	17.93		
meff	9.53	6.22	6.13	6.72	6.00	6.81	11.77	12.19	1.26		

computed from Expression (6). In addition, the design-based test statistic for the coefficient γ_b is:

$$t_{\gamma_b} = \left\{\hat{V}_p(\hat{\gamma}_b)\right\}^{-1/2}\hat{\gamma}_b.$$

Recall that Model ($f1$) has coefficients that are constant across all industries, Model ($f2$) allows different intercept terms across industries and Model ($f3$) allows all coefficients to vary across industries. Also, note that Models ($f1$) through ($f3$) all include both $\ln(x_{j0})$ and $\ln(n_{jt})$. In general, subpopulations with a larger benchmark employment, x_{j0}, will tend to receive larger initial sample sizes and thus also have larger numbers of respondents, n_{jt}, in month t. Consequently, $\ln(x_{j0})$ and $\ln(n_{jt})$ will tend to be positively correlated across our 430 domains j. However, inclusion of both predictors allowed us to account for the effects of the changes in numbers of respondents across months. In Table 3, the positive coefficient on $\ln(n_{jt})$ is an outcome of this positive association between $\ln(x_{j0})$ and $\ln(n_{jt})$. On the other hand, after incorporation of industry-specific intercept terms in Models ($f2$) and ($f3$), the estimated coefficients for $\ln(n_{jt})$ are negative.

In addition, the final rows of Tables 3 through 5 present "misspecification effect" ratios for each of the estimated coefficients. In a slight extension of the ideas in Skinner (1986), define the misspecification effect ratio for the coefficient estimator $\hat{\gamma}_b$ as:

$$\text{meff}_{mb} = \left[\frac{se_{f_m,complex}(\hat{\gamma}_b)}{se_{f_m,direct}(\hat{\gamma}_b)}\right]^2, \tag{29}$$

where $se_{f_m,complex}(\hat{\gamma}_b)$ is the estimated standard error of the ordinary least squares coefficient estimator $\hat{\gamma}_b$ computed with Expression (6) for model f_m; and $se_{f_m,direct}(\hat{\gamma}_b)$ is the corresponding standard error obtained directly from ordinary least squares results, without any adjustment for the correlation across $\hat{V}_{pjt}$ terms induced by the CES design and estimation methods. For cases in which meff_{mb} is greater than one, direct use of unadjusted errors from ordinary least squares regression output will lead to confidence intervals for $\hat{\gamma}_b$ that are too narrow and that have coverage rates below their nominal levels. As one would expect in the analysis of data with relatively strong correlation over time, Table 3 reports misspecification effect ratios that are substantially greater than one for the coefficients γ_0, γ_1 and γ_2. For γ_3 (the coefficient of the $\ln(t)$ predictor), the misspecification effect ratio is close to one. Tables 4 and 5 display qualitatively similar patterns for their misspecification effect ratios, with the exception of the coefficients for Industry 1. This industry had data for only two MSAs, while Industries 2 through 6 had data for 36, 61, 131, 100 and 100 MSAs, respectively.

Table 5. Coefficient estimates and inferential statistics for Model (f3)

	Intercept						$\ln(x_{j0})$							
	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	R^2	$\hat{\sigma}_e^2$
est.	8.04	−3.77	−2.24	−4.30	−2.13	−7.86	0.36	2.00	1.62	1.57	1.32	2.11	0.62	1.04
s.e.	0.46	1.46	1.03	2.28	1.63	1.03	0.15	0.25	0.14	0.36	0.23	0.14		
t_γ	17.50	−2.58	−2.19	−1.89	−1.31	−7.62	2.44	8.03	11.29	4.39	5.66	14.88		
meff	0.02	3.04	2.85	17.57	7.21	3.40	0.05	6.52	5.34	29.05	10.24	5.55		
	$\ln(n_{jt})$						$\ln(t)$							
	γ_{21}	γ_{22}	γ_{23}	γ_{24}	γ_{25}	γ_{26}	γ_{31}	γ_{32}	γ_{33}	γ_{34}	γ_{35}	γ_{36}		
est.	−0.18	−0.97	−0.44	−0.37	−0.13	−1.10	1.42	0.93	0.95	1.31	1.11	1.76		
s.e.	0.45	0.44	0.20	0.46	0.29	0.19	0.22	0.18	0.13	0.13	0.15	0.15		
t_γ	−0.40	−2.22	−2.22	−0.81	−0.43	−5.86	6.47	5.22	7.34	9.85	7.60	11.85		
meff	0.06	7.98	4.34	30.02	9.93	5.22	0.06	0.69	0.62	1.43	1.30	1.32		

Table 6. Wald test of $\omega_0 = \omega_1 = 0$ for Model (24). (Reference value: 6.00 at $\alpha = 0.05$)

First phase model	$f1$	$f2$	$f3$
Test statistics	3.22	2.63	3.13

In applying the residual-analysis methods developed in Section 4 and Appendix C, we used the estimators

$$V^*_{pjt,f_m} = \exp\left(X_{j,f_m}\hat{\gamma}_{f_m} + 2^{-1}\hat{\sigma}^2_{e,f_m}\right)$$

where X_{j,f_m} and $\hat{\gamma}_{f_m}$ are respectively the vectors of predictor variables and ordinary least squares coefficient estimators for a given model m, with each of Models $(f1)$ through $(f3)$ considered separately. In addition, $\hat{\sigma}^2_{e,f_m}$ is the residual mean squared error from the ordinary least squares regression fit for model m. See Karlberg (2000) for related comments.

5.2. *Goodness-of-Fit Measures for the GVF Models*

To evaluate the goodness-of-fit of our GVF models, note first that Tables 3, 4 and 5 present the aggregate measures R^2 equal to 0.52, 0.61 and 0.62 for Models $(f1)$ through $(f3)$, respectively; and the corresponding residual mean squared error terms $\hat{\sigma}^2_e$ are 1.31, 1.06 and 1.04, respectively. Thus, in a summary evaluation of fit across all domains, Model $(f2)$ is somewhat better than $(f1)$, but $(f3)$ is only marginally better than Model $(f2)$. In keeping with the comments following Expression (17), interpretation of R^2 and $\hat{\sigma}^2_e$ values warrants careful consideration of the effect of $V(\epsilon^*_{jt})$. Specifically, applications of the residual-analysis methods from Section 4 indicate several important ways in which Model $(f3)$ may provide a better fit than Models $(f1)$ or $(f2)$ for the CES data.

First, for each of Models $(f1)$ through $(f3)$, Table 6 reports the results of standard Wald test statistics for the null hypothesis $H_0 : \omega_0 = 0 = \omega_1$:

$$W = (\hat{\omega}_0, \hat{\omega}_1)\left[\hat{V}\{(\hat{\omega}_0, \hat{\omega}_1)'\}\right]^{-1}(\hat{\omega}_0, \hat{\omega}_1)',$$

where $\hat{\omega} = \left[\hat{\omega}_0, \quad \hat{\omega}_1, \quad \hat{\omega}_2\right]'$ is computed through an ordinary least squares fit to Model (24) with $\hat{V}(\hat{\omega}_0, \hat{\omega}_1)$ computed as shown in Appendix A. In addition, $\hat{\omega}$ and $\hat{V}(\hat{\omega})$ are based on data from a total of 430 area-industry combinations. Application of the quadratic form ideas reviewed in Appendix A, with $d = 430 - 1 = 429$ and $p = 2$, indicates that $(W/429)\{(429 - 2 + 1)/2\}$ has approximately a noncentral F distribution with 2 and $429 - 2 + 1 = 428$ degrees of freedom and with noncentrality parameter $W_0 = (\omega_0, \omega_1)\left[V\{\hat{\omega}_0, \hat{\omega}_1)'\}\right]^{-1}(\omega_0, \omega_1)'$. In our example, all test statistics from Models

Table 7. Degrees of Freedom (d) among Models (f) given Model (24) with $\omega_0 = \omega_1 = 0$*

Model	$f1$	$f2$	$f3$
ω_2	0.484	0.216	0.004
$se(\omega_2)$	(0.053)	(0.048)	(0.001)
$d*$	4.13	9.25	468.77

($f1$) to ($f3$) were smaller than the reference value, $F_{\{2,428\}}(2)(429)/428 = 6.00$, at $\alpha = 0.05$.

Table 7 reports the estimates $\hat{\omega}_2$ and their standard errors computed under the reduced form of Model (25) with the constraints $\omega_0 = 0 = \omega_1$. Note that Model ($f3$) has large estimated values for $d* = \hat{\omega}_2^{-1}2$, while Models ($f1$) and ($f2$) have much smaller estimated values for $d*$.

Second, we computed the terms r_{jt} from Expression (19) for each of Models ($f1$) through ($f3$) respectively. Figure 1 presents a plot of the resulting r_{jt} against the corresponding predicted values $\ln(V^*_{pjt})$ for Model ($f3$). The grey circles display the plot of r_{jt}, an approximately unbiased estimator of the mean squared error of V^*_{pjt}, against $\ln(V^*_{pjt})$; and the solid black circles display the values of $\hat{h}_{f_3}$, the smoothed version of r_{jt} based on Expression (21) computed for the reduced Model (25). Figure 1 also includes results from a nonparametric regression method known as locally weighted regression (loess) with a span of 0.1. For general background on loess methods, see Cleveland and Grosse (1991). Note that the loess-smoothed estimates are relatively close to the corresponding values of $\hat{h}_{f_3}$ in Figure 1.

Similar plots were produced for Models ($f1$) and ($f2$) but are not shown in the article. For the relatively simple Model ($f1$), the resulting plot indicates that $\hat{h}_{f_1}$, the estimator of $E(q^2_{jt}|X_{jt})$, is relatively large for large values of $\ln(V^*_{jt})$, reflecting a potential lack of fit for Model ($f1$) in this upper range. For ($f2$), which is a more refined model than ($f1$), the corresponding values of $\hat{h}_{f_2}$ are not as large as $\hat{h}_{f_1}$ for high values of $\ln(V^*_{jt})$, indicating a somewhat better fit of ($f2$). In addition, for cases with positive values of r_{jt}, we plotted

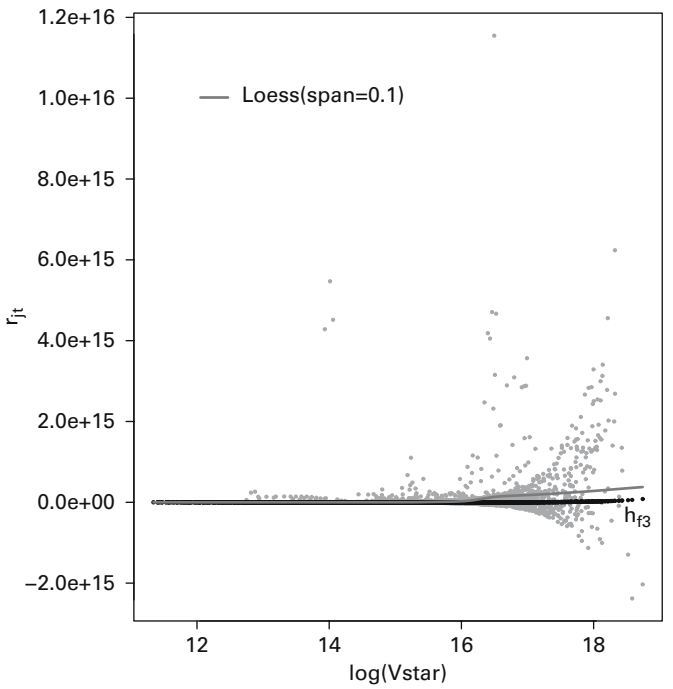

Fig. 1. Three overlaid plots of estimates of $E\left(q^2_{jt}|X_{jt}\right)$ *against* $ln\left(V^*_{pjt}\right)$ *based on Model (*$f3$*). The grey circles present* r_{jt} *based on Expression (19). The grey line presents loess-smoothed values of* r_{jt} *with span* $= 0.1$*. The solid black circles present values of* h_{f3} *computed from the reduced Model (25)*

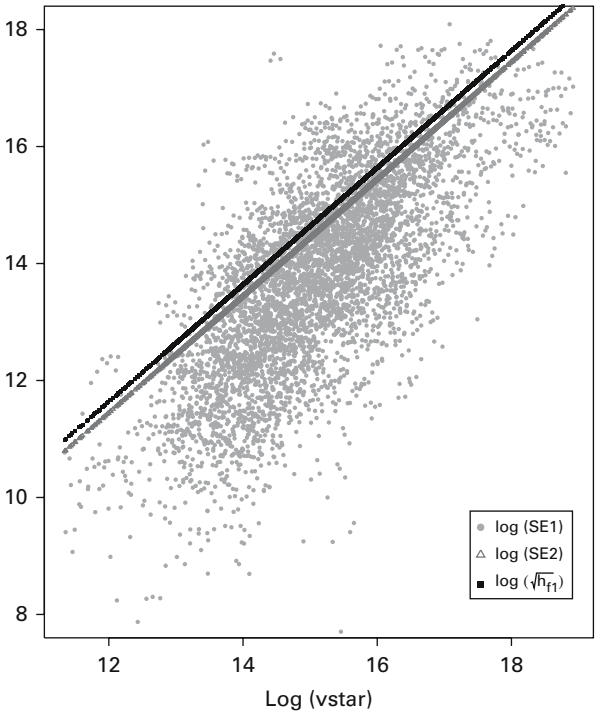

Fig. 2. Plot of ln (SE 1) (grey circles), ln (SE 2) (grey triangles) and ln $(\sqrt{h_{f1}})$ (black squares) against ln $\left(V^{*}_{pjt}\right)$ for the reduced form (25) of the regression model for the error terms r_{jt}. Here, SE 2, $\sqrt{h_{f1}}$ and V^{*}_{pjt} are all based on Model (f 1)

points of $\ln(r_{jt})$ against $\ln(V^{*}_{pjt})$ for Model ($f3$) (again not included here). A loess-smoothed line (span $= 0.1$) drawn through the plotted points was roughly consistent with a linear relationship between $\ln(r_{jt})$ and $\ln(V^{*}_{pjt})$. Furthermore, for all values (j,t), the computed values $\hat{h}_{f_1}$, $\hat{h}_{f_2}$ and $\hat{h}_{f_3}$ were all greater than zero, thus addressing the negative individual values of r_{jt} noted in Subsection 4.4.

Figure 2 plots three measures of uncertainty in prediction of the true design variance V_{pjt}. The first measure, $SE1$, equals the square root of $\left(2\hat{V}^{2}_{pjt}\right)/(d+2)$, which is an unbiased direct estimator of the variance of the prediction error $\hat{V}_{pjt} - V_{pjt}$ under the moment condition (10). The second measure, $SE2$, equals the square root of $\left(2V^{*2}_{pjt}\right)/d$, where V^{*}_{pjt} is computed under Model ($f1$). Under Model ($f1$) and condition (10), $\left(2V^{*2}_{pjt}\right)/d$ is approximately unbiased for the variance of the prediction error $\hat{V}_{pjt} - V_{pjt}$. Thus $SE2$ may be considered as a smoothed version of $SE1$. The third measure, $\sqrt{h_{f_1}}$, is an estimator of the standard deviation of the equation error term q_{jt} under Model ($f1$) and the conditions outlined in Section 4. In Figure 2, the curve for $\ln(\sqrt{h_{f_1}})$ falls slightly above the curve for ln ($SE2$), which indicates that under the relatively simple Model ($f1$), use of the GVF will lead to an estimated standard error for prediction of V_{pjt} that is slightly larger than the standard error of $\hat{V}_{pjt}$ as a predictor of V_{pjt}. Figures 3 and 4 present the corresponding plots of ln ($SE1$), ln ($SE2$) and ln $(\sqrt{h_f})$ against ln (V^{*}_{pjt}) for Models ($f2$) and ($f3$), respectively. Note that in Figure 3, the curve for ln $(\sqrt{h_{f_2}})$ is slightly below the curve for ln ($SE2$), while in Figure 4, ln $(\sqrt{h_{f_3}})$ is substantially below ln ($SE2$).

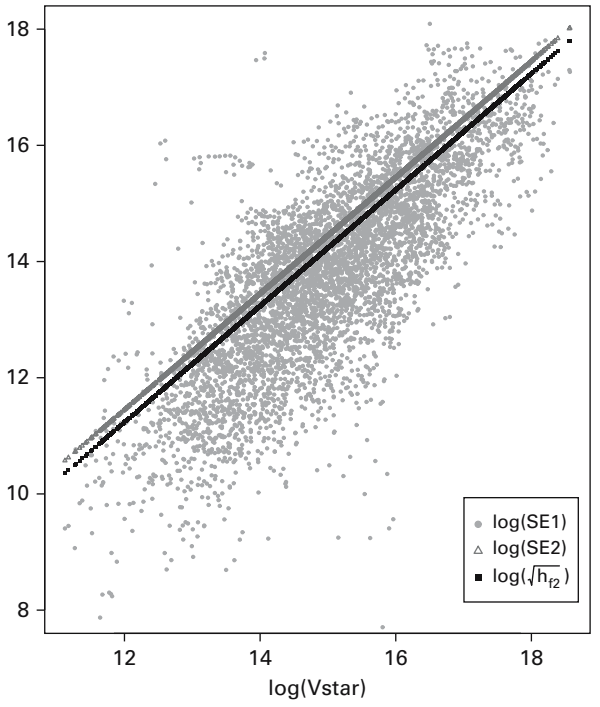

Fig. 3. Plot of ln (SE 1) (grey circles), ln (SE 2) (grey triangles) and ln ($\sqrt{h_{f\,2}}$) (black squares) against ln $\left(V^{*}_{pjt}\right)$ for the reduced form (25) of the regression model for the error terms r_{jt}. Here, SE 2, $\sqrt{h_{f\,2}}$ and V^{*}_{pjt} are all based on Model (f2)

Figure 5 displays plots of $\sqrt{h_f}$ against $\ln(V^{*}_{pjt})$ where both $\sqrt{h_f}$ and $\ln(V^{*}_{pjt})$ are computed separately for each of Models (f1) through (f3). For relatively large values of $\ln(V^{*}_{pjt})$, the curve for (f3) is substantially below the curves for (f1) and (f2). Thus, Figures 2 through 5 indicate that for prediction of the true variances V_{pjt}, under the specified conditions, use of Model (f3) is substantially better than use of either Models (f1) or (f2), or use of the directly computed terms $\hat{V}_{pjt}$. Finally, note that all figures present data for the same area-industry-month combinations from the calendar year 2000. Consequently, some common outlier patterns appear in several of the figures. For example, Figure 1 displays three large positive outliers corresponding to $\ln(V^{*}_{pjt})$ values approximately equal to 14.5. These three points represent three consecutive months for one specific area-industry combination. Similar three-point outlier patterns for the same area-industry combinations appear in Figures 2 through 4.

6. A Simulation Study

6.1. Design of the Study

To evaluate the properties of $\hat{\gamma}$ and V^{*}_{pjt}, we carried out a simulation study based on the following variables produced for each of $R = 1{,}000$ replicates.

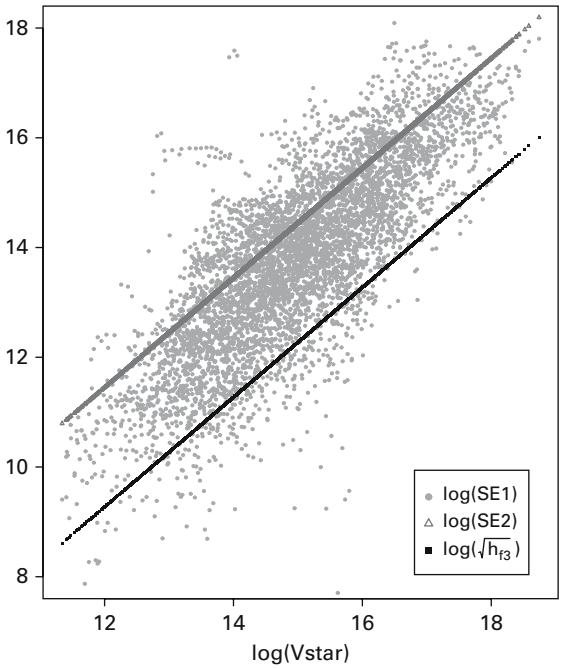

Fig. 4. Plot of ln (SE 1) (grey circles), ln (SE 2) (grey triangles) and ln ($\sqrt{h_{f3}}$) (black squares) against ln$\left(V^{*}_{pjt}\right)$ for the reduced form (25) of the regression model for the error terms r_{jt}. Here, SE 2, $\sqrt{h_{f3}}$ and V^{*}_{pjt} are all based on Model ($f3$)

First, we computed the fixed values

$$f_{1jt} = \gamma_0 + \gamma_1 \ln(x_{j0}) + \gamma_2 \ln(n_{jt}) + \gamma_3 \ln(t) \tag{30}$$

based on the numerical values of the coefficient vector γ for Model ($f1$) presented in Table 3 for all 5,160 combinations of domain j and month t considered in Section 5.

Second, we generated the normal $(0, \sigma^2_{q*})$ random variables $q^{*}_{jt(r)}$ for the 5,160 cases with σ^2_{q*} defined by Expression (C.6) using values of d_q specified in Table 8. We then computed

$$V_{pjt(r)} = \exp(f_{1jt} + q^{*}_{jt(r)}).$$

In addition, we generated $\hat{\theta}_{jt(r)}$ as independent normal (x_{j0}, V_{pjt}) random variables and generated $\epsilon^{*}_{jt(r)}$ as independent normal $(0, \sigma^2_{\epsilon*})$ random variables with $\sigma^2_{\epsilon*}$ defined by Expression (C.5) with $d_\epsilon = 6$. We then computed

$$\hat{V}_{pjt(r)} = V_{pjt(r)} \exp(\epsilon^{*}_{jt(r)}).$$

Based on the 5,160 vectors $\left(\hat{V}_{pjt(r)}, X_{jt}\right)$, where $X_{jt} = \left(1, \ln(x_{j0}), \ln(n_{jt}), \ln(t)\right)$, we carried out ordinary least squares regression of $\ln(\hat{V}_{pjt(r)})$ on X_{jt} to produce the coefficient vector estimate $\hat{\gamma}_{(r)}$; the term $\hat{\sigma}^2_{(r)}$ equal to the regression mean squared error; the term $\hat{\sigma}^2_{q*(r)}$ defined by Expression (C.6); and the predicted variances $V^{**}_{pjt(r)}$ defined by Expression (C.9). In addition, we computed the confidence intervals for θ_{jt},

$$\hat{\theta}_{jt(r)} \pm t_{d_\epsilon, 1-\alpha/2} \left(\hat{V}_{pjt(r)}\right)^{1/2} \tag{31}$$

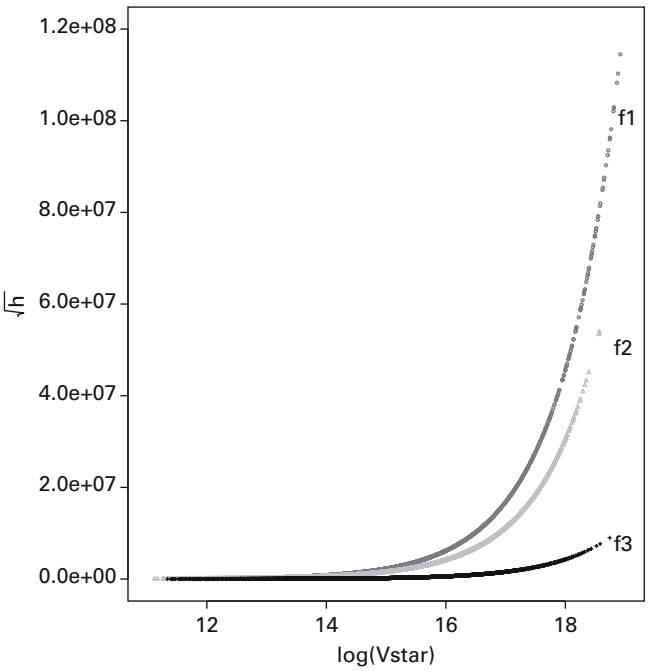

Fig. 5. Three overlaid plots of $\sqrt{h_f}$ against $ln\left(V^{*}_{pjt}\right)$. In the top curve (dark grey circles), both $\sqrt{h_{f1}}$ and $ln\left(V^{*}_{pjt}\right)$ are based on Model (f1) for $ln\left(V^{*}_{pjt}\right)$. In the middle curve (light grey triangles), both $\sqrt{h_{f2}}$ and $ln\left(V^{*}_{pjt}\right)$ are based on Model (f2). In the bottom curve (black crosses), both $\sqrt{h_{f3}}$ and $ln\left(V^{*}_{pjt}\right)$ are based on Model (f3). In all curves, $\sqrt{h_f}$ is based on the reduced Model (25) for r_{jt}.

based on the direct variance estimates $\hat{V}_{pjt(r)}$; and

$$\hat{\theta}_{jt(r)} \pm t_{d_q,1-\alpha/2}\left(V^{**}_{pjt(r)}\right)^{1/2} \tag{32}$$

based on the GVF predictors $V^{**}_{pjt(r)}$, where $t_{d,1-\alpha/2}$ is the upper $1-\alpha/2$ quantile of a t distribution on d degrees of freedom. Finally, taking averages over the R replicates, we computed estimates

$$R^{-1}\sum_{r=1}^{R}\left(\hat{\gamma}_{(r)} - \gamma\right) \tag{33}$$

of the biases of the coefficient estimates;

$$\left(n^{-1}R^{-1}\sum_{r=1}^{R}\sum_{t=1}^{12}\sum_{j=1}^{430}V_{pjt}\right)^{-1}\left(n^{-1}R^{-1}\sum_{r=1}^{R}\sum_{t=1}^{12}\sum_{j=1}^{430}\Delta_{pjt(r)}\right) \tag{34}$$

the aggregate relative bias of the predictors $V^{**}_{pjt(r)}$ where $\Delta_{pjt(r)} = V^{**}_{pjt(r)} - V_{pjt}$, and $n = J \times T = 430 \times 12 = 5,160$;

$$\left(n^{-1}R^{-1}\sum_{r=1}^{R}\sum_{t=1}^{12}\sum_{j=1}^{430}V^{-1}_{pjt}\Delta_{pjt(r)}\right) \tag{35}$$

Table 8. Simulation results for coefficient estimators and variance predictors under Model (f1)

		Coefficient Bias (Standard Deviation)				Var (predictors)		Confidence Interval Properties			
						rel bias aggregated ($*10^{-4}$)	*rel bias domain*	*Cov. Rate*		*Mean Width* ($*10^3$)	
d_q	σ^2_{q*}	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$			$\hat{V}_{pjt}$	V^{**}_{pjt}	$\hat{V}_{pjt}$	V^{**}_{pjt}
4	0.645	0.0010 (0.247)	−0.0001 (0.026)	0.0009 (0.034)	−0.0010 (0.061)	4.068	0.906	0.96	0.98	9.17	10.74
6	0.395	0.0009 (0.215)	−0.0000 (0.022)	0.0006 (0.030)	−0.0008 (0.054)	3.906	0.484	0.96	0.98	8.89	8.89
8	0.284	0.0009 (0.200)	−0.0000 (0.021)	0.0006 (0.028)	−0.0007 (0.050)	3.883	0.328	0.96	0.97	8.77	8.15
16	0.133	0.0008 (0.175)	−0.0000 (0.018)	0.0004 (0.025)	−0.0005 (0.044)	4.002	0.143	0.96	0.96	8.60	7.21
30	0.069	0.0007 (0.164)	−0.0000 (0.017)	0.0002 (0.023)	0.0008 (0.042)	4.188	0.072	0.96	0.96	8.54	6.84
60	0.034	0.0007 (0.158)	−0.0000 (0.017)	0.0002 (0.022)	−0.0003 (0.040)	4.395	0.035	0.96	0.95	8.50	6.64
120	0.017	0.0007 (0.154)	−0.0000 (0.016)	0.0001 (0.022)	−0.0002 (0.039)	4.574	0.017	0.96	0.95	8.48	6.55
400	0.005	0.0006 (0.152)	0.0000 (0.016)	0.0000 (0.021)	−0.0001 (0.038)	4.799	0.006	0.96	0.95	8.47	6.48

the average domain-specific relative bias of V^{**}_{pjt}; and the coverage rates and mean widths for the confidence intervals (31) and (32).

We repeated these steps for the eight values of $d_q = 4, 6, 8, 16, 30, 60, 120$ and 400. Results are displayed in Table 8.

6.2. *Numerical Results*

The first two columns of Table 8 present the selected values of d_q and the corresponding values of σ^2_{q*} based on Expression (C.6). Note that the value $d_q = 4$ corresponds approximately to the value of $d*$ for Model ($f1$) in Table 7; and the value of $d_q = 400$ is slightly less than the value of $d*$ for Model ($f3$) in Table 7.

The next four columns of Table 8 present the bias terms as given in Expression (33), with the corresponding simulated standard deviations placed in parentheses. Note that the bias terms are all small relative to the coefficient values in Table 3 and relative to the reported standard deviations.

The next two columns report the relative bias values given by Expressions (34) and (35), respectively. Note that the aggregate bias terms (34) are relatively small for all cases; while the relative bias terms (35) are fairly large for $d_q = 4$, and decline to values close to zero as d_q increases. The ninth through twelfth columns report coverage rates and mean widths for nominal 95% confidence intervals (31) and (32), respectively. Note that all coverage rates exceed the nominal value of 0.95.

For $d_q = 4$, the intervals (31) based on $\hat{V}_{pjt}$ have a mean width approximately 17% less than the intervals (32) based on V^{*}_{pjt}. This is not surprising, since in this case d_ϵ is greater than d_q. For $d_q = 6$, the intervals (31) and (32) have approximately the same mean width. As d_q increases in the remainder of Table 8, mean widths of the intervals (32) became progressively smaller relative to the widths of the interval (31). This reflects the increasing efficiency of V^{**}_{pjt} relative to $\hat{V}_{pjt}$ as d_q increases with d_ϵ held equal to 6. We observed similar patterns in comparisons of the quantiles of the widths of the confidence intervals (31) and (32); details are omitted here in the interest of space.

In addition, we produced month-specific forms of the final six columns of Table 8, and explored the numerical results for possible time effects. In results not detailed here, we did not identify any substantial time effects for the relative-bias results related to Expressions (34) and (35), nor for the coverage rates of confidence intervals for θ_{jt} based on Expressions (31) and (32), respectively. As one would expect from the positive coefficient γ_3 in Expression (30), the widths of the intervals (31) and (32) did increase over time, but for a given value of d_q, the relative widths of intervals (31) and (32) remained approximately the same.

7. Discussion

7.1. *Summary of Ideas and Methods*

This article has considered two related approaches to the evaluation of generalized variance functions for the analysis of complex survey data. First, an extension of standard estimating equation methods led to design-based variance estimators for the coefficient estimators of a GVF model. This in turn led to design-based inferences for these coefficients, as illustrated by the CES example in Tables 3 through 5. For many of the

coefficients considered in Tables 3 through 5, the numerical values of the misspecification effect ratio (29) were substantially greater than one. Thus, in inference for the CES example, it was important to use the design-based variance estimator from (6) instead of the customary variance estimates obtained directly from standard OLS output. Second, additional conditions on the equation error terms q_{jt} led to approximations for the mean squared error of the GVF-based estimators V^*_{pjt}. A regression model for these MSE terms allowed the comparison of the predictive precision of the GVF V^*_{pjt} with the direct design-based variance estimators $\hat{V}_{pjt}$. Application of this second set of analyses in Tables 6 and 7 and in Figures 1 through 4 allowed the identification of some specific GVFs with smaller MSEs than $\hat{V}_{pjt}$ for our CES data.

7.2. *Possible Extensions*

In closing, we note several possible extensions of the current work. First, we have focused on modeling of the variance of sampling error alone. In some work with small domain estimation, there is also interest in modeling of the variances of prediction errors, which may include components of both sampling error and model error. Second, one may develop additional diagnostics that are specifically focused on evaluation of the effect of GVF lack of fit on specific statistics, that is, confidence intervals for finite population means or variance-based weights in construction of weighted least squares estimators. Third, in keeping with the comments at the end of Subsection 4.4, one could consider estimators of $E(q^2_{jt}|X_{jt})$ based on restricted maximum likelihood methods from the variance component literature. Fourth, Valliant (1987) explored questions regarding use of ordinary least squares or weighted least squares methods in estimation of the coefficients of a GVF model. It would be useful to extend his approach to the context defined in the current article, especially for estimation of the coefficients of the h_f models like (24) and (25). Fifth, the numerical work in this article used the assumption that the equation errors q_{jt} and estimation errors ϵ_{jt} followed lognormal distributions. One could consider extensions of this work to cases in which q_{jt} and ϵ_{jt} follow chi-square distributions or other distributions in the gamma family. Finally, the simulation-based evaluations in Section 6 used values $q^*_{jt(r)}$, $\epsilon^*_{jt(r)}$ and $\hat{\theta}_{jt(r)}$ generated from independent normal distributions. As suggested by a referee, one could carry out related simulation work by expanding the available CES data into a fixed finite population, and then drawing multiple stratified samples from that population.

Appendix A

Development of the Variance Estimator $\hat{V}(\hat{\omega})$

Subsection 3.2 developed variance estimators $\hat{V}(\hat{\gamma})$ for the GVF coefficient estimators $\hat{\gamma}$. To develop a similar estimator for the variance of the approximate distribution of $\hat{\omega}$, define r, $\mathbf{Z}$, J and C as in Subsection 4.5. Under regularity conditions, $\hat{\omega}$ follows approximately a multivariate normal distribution with mean ω and variance-covariance matrix $\mathbf{V}(\hat{\omega})$. An estimator of $\mathbf{V}(\hat{\omega})$ is

$$\hat{V}(\hat{\omega}) = \left\{\hat{\mathbf{w}}^{(1)}(\omega*)\right\}^{-1} \hat{V}\{\hat{\mathbf{w}}(\hat{\omega})\} \left[\left\{\hat{\mathbf{w}}^{(1)}(\omega*)\right\}'\right]^{-1},$$

where $\hat{\mathbf{w}}^{(1)}(\omega*) = \frac{\partial \hat{\mathbf{w}}(\omega)}{\partial \omega}\Big|_{\omega=\omega*} = \mathbf{Z}'\mathbf{Z}$. For Model (20),

$$\hat{\omega} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$
$$\hat{\mathbf{w}}(\hat{\omega}) = \mathbf{Z}'\mathbf{Y} - \mathbf{Z}'\mathbf{Z}\hat{\omega}$$
$$\hat{V}\{\hat{\mathbf{w}}(\hat{\omega})\} = (J-1)^{-1}J\sum_{j\in\mathcal{D}}\{\hat{\mathbf{w}}_{jt}(\hat{\omega}) - \hat{\bar{\mathbf{w}}}(\hat{\omega})\}\{\hat{\mathbf{w}}_{jt}(\hat{\omega}) - \hat{\bar{\mathbf{w}}}(\hat{\omega})\}'$$
$$\text{and} \quad \hat{\bar{\mathbf{w}}}(\hat{\omega}) = J^{-1}\sum_{j\in\mathcal{D}}\hat{\mathbf{w}}_{jt}(\hat{\omega}).$$

Under additional regularity conditions, $d\hat{V}(\hat{\omega})$ follows approximately a Wishart $\big(d, V(\hat{\omega})\big)$ distribution. Standard arguments (e.g., Korn and Graubard 1990) indicate that for a fixed $p \times C$ dimensional matrix $\mathbf{A}$, if we define the quadratic form

$$W = (\mathbf{A}\hat{\omega})'\big[\mathbf{A}\hat{V}(\hat{\omega})\mathbf{A}'\big]^{-1}(\mathbf{A}\hat{\omega}),$$

then $(W/d)\{(d-p+1)/p\}$ has approximately a noncentral F distribution with p and $(d-p+1)$ degrees of freedom and noncentrality parameter $W_0 = (\mathbf{A}\omega)'\big[\mathbf{A}V(\hat{\omega})\mathbf{A}'\big]^{-1}(\mathbf{A}\omega)$.

Appendix B

Ad Hoc "Degrees of Freedom" Measures for Estimation and Prediction Errors Under Variance Function Models

Numerical work in this article uses the assumption that the errors q_{jt} and ϵ_{jt} follow lognormal distributions. However, direct statements about the moments of q_{jt} and ϵ_{jt} may be somewhat difficult to interpret. Consequently, it is useful to provide the following ad hoc "degrees of freedom" measures related to the moments of q_{jt} and ϵ_{jt}.

Let A be a positive random variable with finite positive mean and variance. Then under a standard approach (e.g., Satterthwaite 1941 and Kendall and Stuart 1968, p. 83), the random variable $\{E(A)\}^{-1}dA$ has the same first and second moments as those of a χ^2_d random variable, where we define the "degrees of freedom" term

$$d = \{V(A)\}^{-1}2\{E(A)\}^2. \tag{B.1}$$

Specifically, for the random variables V_{pjt} and $\hat{V}_{pjt}$ defined in Expressions (1) and (2), $\{f(X_{jt},\gamma)\}^{-1}d_{q_{jt}}V_{pjt}$ has the same first and second moments as a $\chi^2_{d_{qjt}}$ random variable, where

$$d_{q_{jt}} = \{V(q_{jt})\}^{-1}\,2\{f(X_{jt},\gamma)\}^2. \tag{B.2}$$

Similarly, conditional on V_{pjt}, $\big(V_{pjt}\big)^{-1}d_{\epsilon_{jt}}\hat{V}_{pjt}$ has the same first and second moments as a $\chi^2_{d_{\epsilon jt}}$ random variable, where

$$d_{\epsilon_{jt}} = \big\{V(\epsilon_{jt}|X_{jt})\big\}^{-1}2\big(V_{pjt}\big)^2. \tag{B.3}$$

Appendix C

Predictors V^*_{jt} of the Design Variance V_{jt} Under Lognormal Models for Equation Error and Estimation Error

Under the model defined by Expressions (2) and (3), define $\epsilon^*_{jt} = \ln(\hat{V}_{jt}) - \ln(V_{jt})$ and assume that

$$q^*_{jt} \sim N(0, \sigma^2_{q*}) \tag{C.1}$$

and

$$\epsilon^*_{jt} \sim N\left(0, \sigma^2_{\epsilon*}\right). \tag{C.2}$$

Then routine calculations show that

$$E(V_{jt}|X_{jt}) = \exp\left(X_{jt}\gamma + 2^{-1}\sigma^2_{q*}\right). \tag{C.3}$$

Let $\hat{\sigma}^2_e$ be the customary mean squared error term from the regression of $\ln(\hat{V}_{pjt})$ on X_{jt} under the model defined by Expressions (2) and (3). Under additional regularity conditions, $\hat{\sigma}^2_e$ is a consistent estimator for the sum $\sigma^2_{q*} + \sigma^2_{\epsilon*}$.
If one does not have satisfactory information about the estimation-error variance term $\sigma^2_{\epsilon*}$, then one may consider use of the predictor

$$V^*_{pjt} = \exp\left(X_{jt}\hat{\gamma} + 2^{-1}\hat{\sigma}^2_e\right). \tag{C.4}$$

Expression (C.4) provides a predictor of the true variance V_{pjt} that is conservative in the sense that $E(V^*_{pjt})$ will tend to be larger than $E(V_{pjt})$. To develop a less conservative predictor of V_{pjt}, suppose that under Expression (B.3), the term $d_{\epsilon_{jt}}$ is known (up to a reasonable level of approximation) and equals the constant d_ϵ for all j and t. Additional calculations for the moments of the lognormal distribution then show that

$$\sigma^2_{\epsilon*} = \Psi\left(1, 2^{-1}d_\epsilon\right) \tag{C.5}$$

where $\Psi(a, b)$ is the Ψ function with arguments a and b (Abramowitz and Stegun 1972, p 258). Similarly, under the lognormal model (C.1), define $d_q = \{V(q_{jt})\}^{-1}2\{E(V_{jt}\}^2$, then

$$\sigma^2_{q*} = \Psi\left(1, 2^{-1}d_q\right) \tag{C.6}$$

In addition, define the function $c(d) = \Psi\left(1, 2^{-1}d\right)$. Expression (C.5) then leads to the estimators

$$\hat{\sigma}^2_{q*} = \hat{\sigma}^2_e - \sigma^2_{\epsilon*} \tag{C.7}$$

and

$$\hat{d}_q = c^{-1}\left(\hat{\sigma}^2_{q*}\right) \tag{C.8}$$

Finally, based on substitution of $\hat{\gamma}$ for γ and $\hat{\sigma}^2_{q*}$ for σ^2_{q*} in Expression (C.3), define the predictor

$$V^{**}_{pjt} = \exp\left(X_{jt}\hat{\gamma} + 2^{-1}\hat{\sigma}^2_{q*}\right). \tag{C.9}$$

8. References

Abramowitz, M. and Stegun, I.A. (1972). Handbook of Mathematical Functions. New York: Dover Publications, INC.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51, 279–292.

Bureau Of Labor Statistics (1997). Employment, Hours, and Earnings from the Establishment Survey, Chapter 2 of BLS Handbook of Methods, U.S. Department of Labor.

Butani, S., Stamas, G., and Brick, M. (1997). Sample Redesign for the Current Employment Statistics Survey. JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 517–522.

Cho, M.J., Eltinge, J.L., Gershunskaya, J., and Huff, L. (2002). Evaluation of the Predictive Precision of Generalized Variance Functions in the Analysis of Complex Survey Data. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 534–539. Available at: http://www.amstat.org/sections/SRMS/Proceedings/ (accessed September 19, 2013).

Cleveland, W.S. and Grosse, E. (1991). Computational Methods for Local Regression. Statistics and Computing, 1, 47–62.

Corbeil, R.R. and Searle, S.R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. Technometrics, 18, 31–38.

Davidian, M., Carroll, R.J., and Smith, W. (1988). Variance Functions and the Minimum Detectable Concentration in Assays. Biometrika, 75, 549–556. DOI: http://www.dx.doi.org/10.1093/biomet/75.3.549

Draper, N.R. and Smith, H. (1998). Applied Regression Analysis, (Third Edition). New York: Wiley.

Eltinge, J.L., Fields, R.C., Gershunskaya, J., Getz, P., Huff, L., Tiller, R., and Waddington, D. (2001). Small Domain Estimation in the Current Employment Statistics Program, Unpublished Background Material for the FESAC Session on Small Domain Estimation at the Bureau of Labor Statistics.

Fuller, W.A. (1987). Measurement Error Models. New York: Wiley.

Gershunskaya, J. and Lahiri, P. (2005). Variance Estimation for Domains in the U.S. Current Employment Statistics Program. In JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 3044–3051. Available at: http://www.amstat.org/sections/SRMS/Proceedings/ (accessed September 19, 2013).

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 71, 320–338. DOI: http://www.dx.doi.org/10.1080/01621459.1977.10480998

Johnson, E.G. and King, B.F. (1987). Generalized Variance Functions for a Complex Sample Survey. Journal of Official Statistics, 3, 235–250.

Judkins, D.R. (1990). Fay's Method for Variance Estimation. Journal of Official Statistics, 6, 223–239.

Karlberg, F. (2000). Survey Estimation for Highly Skewed Population in the Presence of Zeros. Journal of Official Statistics, 16, 229–241.

Kendall, M.G. and Stuart, A. (1968). Advanced Theory of Statistics 3. New York: Hafner Publishing Company.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni *t* Statistics. The American Statistician, 44, 270–276.

O'Malley, A.J. and Zaslavsky, A.M. (2005). Variance-Covariance Functions for Domain Means of Ordinal Survey Items. Survey Methodology, 31, 169–182.

Patterson, H.D. and Thompson, R. (1971). Recovery of Inter-Block Information When Block Sizes Are Unequal. Biometrika, 58, 545–554.

Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin, 2, 110–114.

Skinner, C.J. (1986). Design Effects in Two-Stage Sampling. Journal of Royal Statistical Society, Series B, 48, 89–99.

U.S. Bureau Of Labor Statistics (2006). Employment & Earnings, U.S. Department of Labor, 53(8).

U.S. Bureau Of Labor Statistics (2011). Employment, Hours, and Earnings from the Establishment Survey, Chapter 2 of BLS Handbook of Methods, U.S. Department of Labor, Available at: http://www.bls.gov/opub/hom/pdf/homch2.pdf (accessed September 19, 2013).

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. Journal of American Statistical Association, 82, 499–508.

Werking, G. (1997). Overview of the CES redesign. In JSM Proceedings, the Section on Survey Research Methods: American Statistical Association, 512–516. Available at: http://www.amstat.org/sections/SRMS/Proceedings/ (accessed September 19, 2013).

Wolter, K.M. (2007). Introduction to Variance Estimation (Second Edition). New York: Springer-Verlag.

Received June 2012
Revised June 2013
Accepted August 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 91–105, http://dx.doi.org/10.2478/jos-2014-0005

A Convenient Method of Decomposing the Gini Index by Population Subgroups

Tomson Ogwang[1]

We propose a convenient method of estimating the within-group, between-group, and interaction components of the overall traditional Gini index from the estimated parameters of underlying "trick regression models" involving known forms of heteroscedasticity related to income. Two illustrative examples involving both real and artificial data are provided. The issue of appropriate standard error of the subgroup decomposition is also discussed.

Key words: Subgroup decomposition; Stochastic approach; Gini index; pseudo-Gini.

1. Introduction

Subgroup decomposition of inequality measures entails determining the proportion of observed inequality that is accounted for by the within-group, between-group, and in some cases the interaction component. Analysis of the trends of overall inequality and its components aids policy makers in devising appropriate inequality-reduction strategies. Kanbur (2006) articulates the policy significance of such decompositions and Radaelli (2010) provides a comprehensive survey of the literature on Gini subgroup decomposition.

In the 1960s, Bhattacharya and Mahalanobis (1967) and Rao (1969) decomposed the traditional Gini index by population subgroups into within-group and between-group components. The two-component decomposition strategy is valid if the subgroup income ranges do not overlap, that is, the richest income-receiving unit (individual, household) in the subgroup with a lower mean income class is not richer than the poorest income-receiving unit in any subgroup with a higher mean income. Subsequently, Pyatt (1976), Silber (1989) and Sastry and Kelkar (1994) decomposed the Gini index by population subgroups into within-group, between-group, and interaction (overlapping) components. In the traditional three-component subgroup decomposition, the within-group component is zero when there is no income inequality within each of the subgroups; the between-group component is the value of the Gini index when the income-receiving units in each subgroup receive the subgroup mean income; and the interaction component indicates the degree of income overlap between the subgroups. The three-component approach to Gini subgroup decomposition is more appealing than its two-component counterpart since it also applies when some subgroup income ranges do overlap. Because of this appeal, the

[1] Department of Economics, Brock University, 500 Glenridge Avenue, St. Catharines, Ontario, Canada L2S 3A1. Email: togwang@brocku.ca

Acknowledgment: The author wishes to thank L. Kwong, J.F. Lamarche, the Associate Editor and three anonymous referees for their valuable comments. However, the usual disclaimer applies.

development of convenient ways of conducting three-component decompositions of the Gini index and the search for alternative ways of viewing the three components continue.

Pyatt's, Silber's and Sastry and Kelkar's subgroup decompositions of the Gini index employ matrix formulations that are not easily amenable to empirical implementation using linear regression methods. Yao and Liu (1996) and Yao (1999) proposed convenient ways of conducting these decompositions using spreadsheets without invoking any regressions. The stochastic approach considered by Ogwang (2000, 2004, 2006, 2007) and Giles (2004) provides a simple way of computing the Gini index from the estimated parameters of an underlying regression model with a known form of heteroscedasticity related to income. The purpose of this article is to exploit the simplification provided by the stochastic approach for purposes of conducting three-component subgroup decomposition of the traditional Gini index.

The methodology considered in this article has three major advantages. First, it provides a new way of viewing the within-group, between-group and interaction components of the traditional Gini index. For example, the interpretation of the between-group and interaction components of the overall Gini as weighted averages of their respective pseudo-Ginis has not previously been featured in the literature on subgroup decomposition of the Gini index. The concept "pseudo-Gini" as used in the context of Gini subgroup decomposition is explained in Section 3, where its interpretation and its similarity to the same concept as used in the context of income source decompositions are also explained. Second, from a practical perspective, the proposed method provides a new and convenient way of computing the three aforementioned components using widely available regression software packages. Third, the calculations and decomposition processes involved are quite transparent, which facilitates understanding.

The format of the rest of the article is as follows. In Section 2 we provide a brief overview of the stochastic approach to the overall Gini index. In Section 3 we derive the salient results that are needed in order to extend the stochastic approach for purposes of subgroup decomposition of the Gini index. In Section 4 we describe the actual empirical implementation of the stochastic approach for Gini subgroup decomposition. The salient issues surrounding the estimation of appropriate standard errors of the Gini subgroup decompositions are discussed in Section 5. Section 6 provides two illustrative examples using both real and artificial data and Section 7 forms the conclusion.

2. The Stochastic Approach to the Gini Index

Let $y_1, y_2, \ldots, y_n$ be the individual incomes of n income-receiving units (individuals, households) which are arranged such that $y_1 \leq y_2 \leq \ldots \leq y_n$, and hence the ranks of y_1 and y_n are 1 and n, respectively. Tied incomes are assigned the average of the ranks they would get without ties.

For purposes of subgroup decomposition of the Gini index using the stochastic approach, it is convenient to utilize the following formula considered by Ogwang (2007)

$$\hat{G} = (1/n)\hat{\gamma} \tag{1}$$

where $\hat{\gamma}$ is the weighted least squares (WLS) estimator of γ in the "trick regression model" $i* = \gamma + u_i$ where $i* = (2i - n - 1), \quad i = 1, 2, \ldots, n$, assuming that the errors u_i are

heteroscedastic of the form $E(u_i^2) = \sigma^2/y_i$ (equivalently, $\hat{\gamma}$ is the ordinary least squares (OLS) estimator of γ in the transformed model $i^*\sqrt{y_i} = \gamma\sqrt{y_i} + u_i\sqrt{y_i}$ where the transformed errors $u_i\sqrt{y_i}$ are homoscedastic under the assumed heteroscedastic structure). Note that under the stipulated heteroscedastic structure, Equation (1) yields accurate point estimates of the Gini index regardless of the number of observations. This is because under this heteroscedastic structure, the WLS estimator of γ in Equation (1) divided by n is, in fact, the usual Gini statistic.

3. Extending the Stochastic Approach to Gini Subgroup Decomposition

Suppose that n income-receiving units are classified into k mutually exclusive and exhaustive subgroups by, for example, gender, age, race, education, occupation or region. The k subgroups are arranged in ascending order of their subgroup mean incomes but the incomes in each subgroup can be in any order. Subgroups with identical mean incomes are first merged into one subgroup.

To facilitate the exposition of subgroup decomposition of the Gini index, let $n_j,\ j = 1, 2, \ldots, k$ denote the number of income-receiving units in the subgroup with the jth smallest mean income, in which case $n = \sum_{j=1}^{k} n_j$. Let $y_{ij}, i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k$ denote the income of the ith income-receiving unit in the subgroup with the jth smallest mean income and $\bar{y}_j = (1/n_j)\sum_{i=1}^{n_j} y_{ij}$, the mean income in the same subgroup.

With respect to the income ranks, let $r_{ij},\ i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k$ denote the rank of y_{ij} in relation to the incomes of all the $n = \sum_{j=1}^{k} n_j$ income-receiving units in all the k subgroups. Also, let $r'_{ij},\ i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k$ denote the rank of y_{ij} in relation to the incomes of only the n_j income-receiving units in the subgroup with the jth smallest mean income. Finally, let $\tilde{r_{ij}} (j = 1, 2, \ldots, k; i = 1, 2, \ldots, n_j)$ denote the rank of y_{ij} in relation to the incomes of all the $n = \sum_{j=1}^{k} n_j$ income-receiving units, assuming that $y_{ij} = \bar{y}_j$ (i.e., all the income-receiving units in each subgroup are assumed to receive the mean income for that subgroup).

To facilitate the exposition of subgroup decomposition of the Gini index by exploiting Equation (1), it is also necessary to define the following three rank vectors: $r_{ij}^* = 2r_{ij} - n - 1 (i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k)$, $r_{ij}^{*\prime} = 2r'_{ij} - n_j - 1\ \ (i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k)$ and $\tilde{r}_{ij}^* = 2\tilde{r}_{ij} - n - 1 (i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k)$. It is easy to verify that

$$r_{ij}^* = \left[2r'_{ij} - n_j - 1\right] + 2\left(r_{ij} - r'_{ij}\right) + (n_j - n),\ \ i = 1, 2, \ldots, n_j;\ \ j = 1, 2, \ldots, k. \qquad (2)$$

Equation (2) can be conveniently rewritten as

$$r_{ij}^* = r_{ij}^{*\prime} + 2(r_{ij} - r'_{ij}) + (n_j - n),\ \ i = 1, 2, \ldots, n_j;\ \ j = 1, 2, \ldots, k. \qquad (3)$$

Since the subgroups are arranged in ascending order of their mean incomes and in the absence of any subgroup mean income ties

$$\tilde{r}_{i1} = (n_1 + 1)/2, \;\; i = 1, 2, \ldots, n_1; \;\; \tilde{r}_{ij} = \left[\sum_{i=1}^{j-1} n_i\right] + (n_j + 1)/2, \tag{4}$$
$$i = 1, 2, \ldots, n_j; \;\; j = 2, \ldots, k$$

(i.e., $\tilde{r}_{i1} = (n_1 + 1)/2$; $\tilde{r}_{i2} = n_1 + (n_2 + 1)/2$; $\tilde{r}_{i3} = n_1 + n_2 + (n_3 + 1)/2$; etc.).

Also,

$$\tilde{r}^*_{i1} = (n_1 - n), \;\; i = 1, 2, \ldots, n_1; \;\; \tilde{r}^*_{ij} = \left[\sum_{i=1}^{j-1} 2n_i\right] + (n_j - 1), \tag{5}$$
$$i = 1, 2, \ldots, n_j; \;\; j = 2, \ldots, k$$

(i.e., $\tilde{r}^*_{i1} = (n_1 - n)$; $\tilde{r}^*_{i2} = 2n_1 + (n_2 - n)$; $\tilde{r}^*_{i3} = 2n_1 + 2n_2 + (n_3 - n)$; etc.).

Substituting $(n_1 - n) = \tilde{r}^*_{i1}, i = 1, 2, \ldots, n_1; (n_j - n) = \tilde{r}^*_{ij} - \sum_{i=1}^{j-1} 2n_i, \; i = 1, 2, \ldots, n_j;$ $j = 2, \ldots, k$ from Equation (5) into Equation (3) and rearranging the terms yields

$$r^*_{ij} = r^{*\prime}_{ij} + \tilde{r}^*_{ij} + 2\widehat{r}^*_{ij}, i = 1, 2, \ldots, n_j; \; j = 1, 2, \ldots, k \tag{6}$$

where $\widehat{r}^*_{i1} = (r_{i1} - r'_{i1}), \; i = 1, 2, \ldots, n_1$ and $\widehat{r}^*_{ij} = \left(r_{ij} - r'_{ij} - \sum_{i=1}^{j-1} 2n_i\right), \; i = 1, 2, \ldots, n_j;$ $j = 2, \ldots, k$. The overall Gini index is given by

$$\hat{G} = \frac{1}{n} \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{k} r^*_{ij} y_{ij}}{\sum_{i=1}^{n_j} \sum_{j=1}^{k} y_{ij}} \tag{7}$$

Substituting r^*_{ij}, given by Equation (6), into Equation (7) yields

$$\hat{G} = \frac{1}{n} \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{k} r^{*\prime}_{ij} y_{ij}}{\sum_{i=1}^{n_j} \sum_{j=1}^{k} y_{ij}} + \frac{1}{n} \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{k} \tilde{r}^*_{ij} y_{ij}}{\sum_{i=1}^{n_j} \sum_{j=1}^{k} y_{ij}} + \frac{2}{n} \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{k} \widehat{r}^*_{ij} y_{ij}}{\sum_{i=1}^{n_j} \sum_{j=1}^{k} y_{ij}}. \tag{8}$$

The first, second, and third terms on the right-hand side of Equation (8) are the within-group component ($\hat{G}_W$), the between-group component ($\hat{G}_B$), and the interaction component ($\hat{G}_I$), respectively, of the Gini index. It turns out that the within-group, between-group and interaction components are weighted averages of the within-group Ginis, the between-group pseudo-Ginis, and the interaction pseudo-Ginis, respectively.

The concept "pseudo-Gini" as used in this article refers to a numerical quantity which is computed using a Gini-like formula as an intermediate step in computing the between-group component of the overall Gini (in the case of the between-group pseudo-Gini) or the interaction component of the overall Gini (in the case of the interaction pseudo-Gini). Fei et al. (1978) and Shorrocks (1982) also use the concept "pseudo-Gini" in a similar manner, but in the context of income-source decompositions of the Gini index. As will be seen below, the subgroup decomposition pseudo-Ginis pertain to the various population

subgroups whereas the income-source decomposition pseudo-Ginis pertain to the various income sources.

For the within-group and between-group components in Equation (8), the weights are equal to the product of the subgroup population shares and the corresponding income shares; for the interaction component, the weights are equal to twice the product of the subgroup population shares and the corresponding income shares. To see these results, we note that the within-group component can be written as

$$\hat{G}_W = \sum_{j=1}^{k} p_j s_j \hat{G}_{Wj}, \;\; j = 1, 2, \ldots, k \tag{9}$$

where $p_j = n_j/n$ is the population share of group j; $s_j = \left(\sum_{i=1}^{n_j} y_{ij}\right) / \sum_{j=1}^{n_j} \sum_{j=1}^{k} y_{ij}$ is the income share of group j; and

$$\hat{G}_{Wj} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} r_{ij}^{*\prime} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}}, \;\; j = 1, 2, \ldots, k$$

is the within-group Gini for group j; the between-group component can be written as

$$\hat{G}_B = \sum_{j=1}^{k} p_j s_j \hat{G}_{Bj}, \;\; j = 1, 2, \ldots, k \tag{10}$$

where p_j and s_j are defined in Equation (9) and

$$\hat{G}_{Bj} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} \tilde{r}_{ij}^{*} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}}, \;\; j = 1, 2, \ldots, k$$

is the between-group pseudo-Gini for subgroup j. Note that $0 \leq \hat{G}_B \leq 1$ even though some between-group pseudo-Ginis may be negative as explained below; and the interaction component can be written as

$$\hat{G}_I = 2 \sum_{j=1}^{k} p_j s_j \hat{G}_{Ij}, \;\; j = 1, 2, \ldots, k \tag{11}$$

where p_j and s_j are defined in Equation (9) and

$$\hat{G}_{ij} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} \widehat{r}_{ij}^{*} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}}, \quad j = 1, 2, \ldots, k$$

is the interaction pseudo-Gini for group j. Note also that $0 \leq \hat{G}_1 \leq 1$ even though some interaction pseudo-Ginis may be negative as explained below.

Before discussing the empirical implementation of the proposed method, it behooves us to clarify the interpretation of the between-group and interaction pseudo-Ginis as defined in Equations (10) and (11), respectively. Since the income shares and population shares in Equations (10) and (11) cannot be negative, and $0 \leq \hat{G}_B, \hat{G}_I \leq 1$, a positive/negative between-group pseudo-Gini for a particular subgroup indicates that the subgroup makes a positive/negative contribution to the between-group component. Likewise, a positive/

negative interaction pseudo-Gini for a particular subgroup indicates that the subgroup makes a positive/negative contribution to the interaction component. It should be noted that the pseudo-Ginis for the income source decompositions, as discussed by Fei et al. (1978) and Shorrocks (1982) among others, like those for the subgroup decompositions, can be positive or negative. With respect to the income source Gini decompositions, a positive/negative pseudo-Gini for a particular income source indicates that the source makes a positive/negative contribution to overall income inequality.

The issue of why the between-group pseudo-Gini or its interaction counterpart may be negative also deserves an explanation. A careful inspection of Equation (9) reveals that the transformed ranks, $r_{ij}^{*\prime}$, $j = 1, 2, \ldots, n_j$, which are used in the computation of $\hat{G}_{wj}$, are identical to the transformed ranks which are used in the calculation of the Gini index for the jth subgroup. Hence, $\hat{G}_{wj}$, $j = 1, 2, \ldots, n_j$, must lie between 0 and 1. Since the transformed ranks, $\tilde{r}_{ij}^{*}$, $j = 1, 2, \ldots, n_j$, in Equation (10), which are used in the computation of $\hat{G}_{Bj}$, are different from the transformed ranks used in the calculation of the Gini index for the jth subgroup, there is no guarantee that $\hat{G}_{Bj}$ will lie between 0 and 1 even though a Gini-like formula is used in its computation. Likewise, since the transformed ranks, $\widehat{r}_{ij}^{*}$ in Equation (11), which are used in the computation of $\hat{G}_{lj}$, are different from the transformed ranks used in the calculation of the Gini index for the jth subgroup, there is also no guarantee that $\hat{G}_{lj}$ will lie between 0 and 1 even though a Gini-like formula is used in its computation. In the illustrative example provided below, negative between-group and interaction pseudo-Ginis are obtained.

4. Empirical Implementation

First we create the seven rank vectors r_{ij}, $r_{ij}{}'$, $\tilde{r}_{ij}$, $r_{ij}{}^{*}$, $r_{ij}{}^{*\prime}$, $\tilde{r}_{ij}{}^{*}$ and $\widehat{r}_{ij}{}^{*}$.

4.1. The Overall Gini Index

An inspection of Equation (7) reveals that the overall Gini index is given by

$$\hat{G} = (1/n)\hat{\gamma} \tag{12}$$

where $\hat{\gamma}$ is the WLS estimator of γ in the model $r_{ij}^{*} = \gamma + u_{ij}^{*}$, $i = 1, 2, \ldots, n_j$; $j = 1,2,\ldots,k$, assuming that the errors, u_{ij}^{*}, are heteroscedastic of the form $E\left(u_{ij}^{*2}\right) = \sigma^2/y_{ij}$.

4.2. The Within-Group Component

An inspection of the first expression on the right-hand side of Equation (8) reveals that the within-group component is given by

$$\hat{G}_w = (1/n)\hat{\gamma}_w \tag{13}$$

where $\hat{\gamma}_w$ is the WLS estimator of γ_w is the model $r_{ij}^{*\prime} = \gamma_w + u_{ij}^{*\prime}$, $i = 1, 2, \ldots, n_j$; $j = 1,2,\ldots,k$, assuming that the errors, $u_{ij}^{*\prime}$, are heteroscedastic of the form $E\left(u_{ij}^{*\prime 2}\right) = \sigma^2/y_{ij}$.

4.3. The Between-Group Component

An inspection of the second expression on the right-hand side of Equation (8) reveals that the between-group component is given by

$$\hat{G}_B = (1/n)\hat{\gamma}_B \tag{14}$$

where $\hat{\gamma}_B$ is the WLS estimator of γ_B in the model $\tilde{r}^*_{ij} = \gamma_B + \tilde{u}^*_{ij}$, $i = 1, 2, \ldots, n_j$; $j = 1, 2, \ldots, k$, assuming that the errors, $\tilde{u}^*_{ij}$, are heteroscedastic of the form $E\left(\tilde{u}^{*2}_{ij}\right) = \sigma^2/y_{ij}$.

4.4. The Interaction Component

An inspection of the third expression on the right-hand side of Equation (8) reveals that the interaction component is given by

$$\hat{G}_I = (2/n)\hat{\gamma}_I \tag{15}$$

where $\hat{\gamma}_I$ is the WLS estimator of γ_I in the model $\hat{r}^*_{ij} = \gamma_I + \hat{u}^*_{ij}$, $i = 1, 2, \ldots, n_j$; $j = 1, 2, \ldots, k$, assuming that the errors, $\hat{u}^*_{ij}$, are heteroscedastic of the form $E\left(\hat{u}^{*2}_{ij}\right) = \sigma^2/y_{ij}$.

Two points about these subgroup decompositions are noteworthy. First, $\hat{\gamma} = \hat{\gamma}_w + \hat{\gamma}_B + 2\hat{\gamma}_I$ and $\hat{G} = \hat{G}_W + \hat{G}_B + \hat{G}_I$. Second, the heteroscedastic structures for the regressions underlying $\hat{G}$, $\hat{G}_W$, $\hat{G}_B$ and $\hat{G}_I$ are similar.

5. Standard Errors of the Gini Subgroup Decompositions

When reporting estimates of the inequality measures and their decompositions, it is also important to report estimates of their standard errors or confidence intervals to facilitate hypotheses tests about their significance. The case for reporting standard errors of inequality measures as well as their decompositions seems strong given that large standard errors may arise even though the number of income-receiving units is large, as pointed out by Maasoumi (1994).

In the case of the overall Gini, four types of standard errors have been employed in the literature, namely the asymptotic standard errors (e.g., Cowell 1989; Davidson 2009), the bootstrap (e.g., Dixon et al. 1987; Mills and Zandvakili 1997; Davidson 2009), the jackknife (e.g., Yitzhaki 1991; Karoly 1992; Ogwang 2000), and WLS/OLS (e.g., Giles 2004). However, most Gini subgroup decomposition proposals so far do not consider the issue of appropriate standard errors of the relevant components, which is surprising given that the standard errors or confidence intervals of the decompositions of other inequality measures, such as the Generalized Entropy class of measures, are widely reported in the literature. For example, Mills and Zandvakili (1997), Biewen (2002), and Gray et al. (2003) report bootstrap standard errors of the subgroup decompositions of several inequality measures, but they do not report the standard errors of the decompositions of the Gini index. Mussard and Richard (2012) is a rare paper that reports confidence intervals of the Gini decompositions.

Although OLS/WLS standard errors of the decompositions could also be obtained from the estimated standard errors of the parameters of the underlying regressions following

Giles (2004), we do not recommend doing so in light of the inadequacies associated with using OLS/WLS-based Gini standard errors raised by Modarres and Gastwirth (2006), Ogwang (2004, 2006); Davidson (2009) and Langel and Tillé (2013), among others, in the context of the overall Gini. These inadequacies arise because OLS/WLS standard errors are based on ordered observations that are not independent even when the income series is independent identically distributed. Although ordering does not affect point estimates of the decompositions, it affects the OLS/WLS-based standard errors. To circumvent this problem, the use of resampling methods (e.g., the bootstrap or jackknife) in conjunction with the stochastic approach to obtain the standard errors of the decompositions is recommended, provided that these methods are applied correctly. For example, Davidson (2009) has noted, in the context of the overall Gini, that bootstrap standard errors can be reliable if applied correctly. More recently, Langel and Tillé (2013) demonstrated that jackknife standard errors of the Gini index can also be reliable if the randomness of the income ranks is properly taken into account by recalculating the ranks each time an observation is dropped in the computation of the jackknife standard error.

Shao and Tu (1995) provide a comprehensive overview of the bootstrap and jackknife techniques in general. Davidson (2009) describes the bootstrap and jackknife approaches to estimating the standard errors and confidence intervals of the overall Gini index. Since that jackknife confidence intervals of the subgroup decompositions of the Gini index are reported in one of the illustrative examples below, we provide a brief description of the jackknife approach to the computation of the standard errors of the subgroup decompositions. Let $\hat{G}_W(n,k)$, $\hat{G}_B(n,k)$ and $\hat{G}_I(n,k)$ denote the estimates of the within-group, between-group and interaction components, respectively, of the Gini index based on the remaining $(n-1)$ observations after deleting the kth observation. Also, let $\bar{G}_W(n) = n^{-1}\sum_{k=1}^{n}\hat{G}_W(n,k)$, $\bar{G}_B(n) = n^{-1}\sum_{k=1}^{n}\hat{G}_B(n,k)$ and $\bar{G}_I(n) = n^{-1}\sum_{k=1}^{n}\hat{G}_I(n,k)$ be the means of all the $\hat{G}_W(n,k), k = 1,2,\ldots,n$, $\hat{G}_B(n,k), k = 1,2,\ldots,n$, and $\hat{G}_I(n,k), k = 1,2,\ldots,n$, respectively. The jackknife standard errors of the within-group, between-group and interaction components of the Gini index are given by

$$SE(\hat{G}_W) = \sqrt{\left(\frac{n-1}{n}\right)\sum_{k=1}^{n}\left(\hat{G}_W(n,k) - \bar{G}_W(n)\right)^2},$$

$$SE(\hat{G}_B) = \sqrt{\left(\frac{n-1}{n}\right)\sum_{k=1}^{n}\left(\hat{G}_B(n,k) - \bar{G}_B(n)\right)^2} \quad \text{and}$$

$$SE(\hat{G}_I) = \sqrt{\left(\frac{n-1}{n}\right)\sum_{k=1}^{n}\left(\hat{G}_I(n,k) - \bar{G}_I(n)\right)^2}, \quad \text{respectively.}$$

6. Illustrative Examples

Example 1: Artificial Data

We first utilize a simple dataset comprising the incomes of $n = 7$ individuals which are broken down into three subgroups with different mean incomes (i.e., $k = 3$) to illustrate

the steps involved in the empirical implementation of the proposed method. Silber (1989) and Sastry and Kelkar (1994) used the same data to demonstrate their Gini subgroup decomposition methods. The incomes constituting the first, second, and third subgroups are {(1,3); mean income = 2}, {(1,4,7); mean income = 4}, and {(6,10); mean income = 8}, respectively. As mentioned above, the subgroups are arranged in ascending order of their mean incomes, in which case $n_1 = n_3 = 2$ and $n_2 = 3$.

Table 1 shows the required ranks and their transformations, using the computational procedures and notations described in Sections 3 and 4. Non-zero entries for $\widehat{r}_{ij}^{*}$ in the last column of the table indicate the existence of a non-zero interaction component since there is an overlap in the income ranges for the first and second subgroups. Table 2 summarizes the empirical results. It is apparent from the table that, apart from rounding errors, $\hat{\gamma} = \hat{\gamma}_w + \hat{\gamma}_B + 2\hat{\gamma}_I$ and $\hat{G} = \hat{G}_W + \hat{G}_B + \hat{G}_I$.

Table 3 compares our decomposition results with previous results based on the same dataset. The table indicates that our decomposition results are identical to Silber's (1989) results. However, Sastry and Kelkar's (1994) estimates of the between-group and interaction components differ from ours. The difference between Sastry and Kelkar's results and our results arises because Sastry and Kelkar assign the income-receiving units their original ranks in the computation of the between-group component.

According to our results, between-group inequality accounts for most of the observed inequality and the interaction term accounts for the smallest percentage of total inequality. Another notable aspect of Table 3 is the numerical equivalence between our subgroup decomposition and Dagum's (1997) decomposition. Specifically, our between-group and interaction components are numerically the same as Dagum's *net contribution between population subgroups* (G_{nb}) and *the contribution of the income intensity of transvariation between subgroups* (G_t), respectively. Hence, Dagum's approach to Gini subgroup decomposition only differs from the traditional approach in the interpretation of some components.

Example 2: Real Data

The methodology described above is also applied to the data on the total pre-tax post-transfer incomes, in Canadian dollars, of a random sample of 4,883 persons, derived from the Canadian Census 2006 Public Use Micro data Files.

*Table 1. Computing the required ranks and their transformations**

Group (i)	y_{ij}	r_{ij}	r'_{ij}	$\tilde{r}_{ij}$	r^{*}_{ij}	$r^{*\prime}_{ij}$	$\tilde{r}^{*}_{ij}$	$\widehat{r}^{*}_{ij}$
1	1	1.5	1	1.5	−5	−1	−5	0.5
1	3	3	2	1.5	−2	1	−5	1
2	1	1.5	1	4	−5	−2	0	−1.5
2	4	4	2	4	0	0	0	0
2	7	6	3	4	4	2	0	1
3	6	5	1	6.5	2	−1	5	−1
3	10	7	2	6.5	6	1	5	0

*$n_1 = n_3 = 2; n_2 = 3; n = 7; k = 3; r^{*}_{ij} = 2r_{ij} - n - 1; r^{*\prime}_{ij} = 2r'_{ij} - n_j - 1;\ \tilde{r}^{*}_{ij} = 2\tilde{r}_{ij} - n - 1; \widehat{r}^{*}_{i1} = (r_{i1} - r'_{i1})$, $i = 1, n_1 = 2$ and $\widehat{r}^{*}_{ij} = \left(r_{ij} - r'_{ij} - \sum_{i=1}^{j-1} n_i\right), i = 1, 2, \ldots, n_j; j = 2,\ k = 3.$

*Table 2. Computing the overall Gini index and its decompositions**

Gini	Equation	$\hat{\gamma}$	$\hat{\gamma}_W$	$\hat{\gamma}_B$	$\hat{\gamma}_I$	$\hat{G}$	$\hat{G}_W$	$\hat{G}_B$	$\hat{G}_I$
Overall	12	2.625				0.375			
Within	13		0.563				0.080		
Between	14			1.875				0.268	
Interaction	15				0.094				0.027

* $n = 7; \hat{G} = \hat{\gamma}/n; \hat{G}_W = \hat{\gamma}_W/n; \hat{G}_B = \hat{\gamma}_B/n$ and $\hat{G}_I = 2\hat{\gamma}_I/n$.

The fact that Canada has one of the highest per capita immigration rates in the world provides a strong case for studying the income differentials between immigrants and nonimmigrants in Canada. One of the common goals of such studies is to examine the factors that influence the intertemporal changes in income inequality between immigrants and nonimmigrants.

In this illustrative example, we consider a breakdown of the data into three nonoverlapping subgroups by immigration status (i.e., nonpermanent residents, nonimmigrants and immigrants). According to the 2006 Census definitions, nonimmigrants are people who are Canadian citizens by birth; immigrants are people who were, or had ever been, landed immigrants in Canada prior to the day of the 2006 Census; and nonpermanent residents are people from other countries who, at the time of the 2006 Census, held a work/study permit or claimed refugee status, as well as their family members living in Canada.

To provide some insights into the nature of the income distributions pertaining to the various subgroups, in Table 4, we report some relevant descriptive statistics pertaining to these subgroups and the full sample. The table indicates positive skewness of the distributions for all income ranges, as would be expected. Another notable feature of Table 4 is the significant overlap in the income ranges. Specifically, the income range [0,110000] for the nonpermanent resident subgroup, with the lowest mean income, lies entirely within the income range [0, 866340] for the immigrant subgroup, with the second lowest mean income, which in turn lies entirely within the income range [0,1285600] for the nonimmigrant subgroup, with the highest mean income. As mentioned above, the interaction component of the Gini index is non-zero if there are overlaps in the income ranges of some of the population subgroups, which is obviously the case in the present example. In fact, given the large overlaps in the income ranges under consideration, we would also expect the contribution of the interaction component to the overall Gini index to be large and statistically significant.

Table 3. Comparing decomposition methods

Method	Overall	Within group	Between group	Interaction
Silber (1989)	0.375	0.080	0.268	0.027
Sastry and Kelkar (1994)	0.375	0.080	0.205	0.089
Dagum (1997)	0.375	0.080	0.268*	0.027**
Present proposal	0.375	0.080	0.268	0.027

*and ** denote estimates of Dagum's (1997) G_{nb} and G_t, respectively.

*Table 4. Income descriptive statistics by population subgroups (Canadian 2006 Census data)**

Subgroup	Mean	Median	Min	Max	Skewness	Excess Kurtosis
Nonpermanent residents	26,487	13,000	0	110,000	1.1641	0.21446
Immigrants	33,338	21,000	0	866,340	7.4692	82.00
Nonimmigrants	34,410	25,000	0	1,285,600	8.6455	180.67
All groups	34,102	24,000	0	1,285,600	8.3156	147.43

* The incomes are measured in Canadian dollars.

In order to gauge the extent of the overlap, we applied the computational procedures described in Sections 3 and 4 above to the 2006 Census data, the results of which are reported in Table 5. Several features of the table are particularly noteworthy. First, the estimate of the overall Gini index turns out to be 0.520 and is statistically significant at the conventional five percent level of significance, since the estimated 95 percent confidence interval does not include zero. Second, the estimates of the within-group, between-group and interaction components turn out to be 0.329, 0.007 and 0.184, respectively. However, only the within-group and interaction component estimates turn out to be statistically significant at five percent level of significance. The confidence interval for the between-group component turns out not to show statistical significance at five percent level of significance, since the associated 95 percent confidence interval includes zero.

It is also apparent from Table 5 that within-group inequality accounts for most of the observed inequality, followed by the interaction component, which is in turn followed by the between-group component. Owing to the huge overlap in the income ranges for the three subgroups, as mentioned above, it is not surprising that the contribution of the interaction component is larger than that of the between-group component. Clearly, these results are informative with respect to the nature of the observed income inequality and are consistent with the descriptive statistics reported in Table 4.

Another positive aspect of the approach adopted in this article is its ability to provide a more detailed breakdown of the subgroup decompositions to also include the contributions of the various subgroups to overall inequality as well as their standard errors. The issue of the contributions of the various subgroups to overall inequality has so far been largely ignored in the literature on Gini subgroup decomposition, which focuses more on the

Table 5. Gini subgroup decomposition results (Canadian 2006 Census data)

Component	Point estimate	95 percent confidence interval*
Within-group	0.3288	(0.3134 to 0.3442)
Between-group	0.0073	(−0.0140 to 0.0285)
Interaction	0.1837	(0.1509 to 0.2165)
Overall	0.5198	(0.4998 to 0.5398)

* The confidence intervals were constructed using the computed jackknife standard errors. To conserve space, the results of the intermediate computational procedures as described in Sections 3, 4 and 5 are not reported but are available from the author on request.

Table 6. A detailed breakdown of the Gini subgroup decompositions (Canadian 2006 Census data)

		Between-group and interaction pseudo-Ginis				Contributions of the various components to overall inequality			
Immigration status (j)	Within-group Gini $\hat{G}_{Wj}$	$\hat{G}_{Bj}$	$\hat{G}_{Ij}$	Pop. share p_j	Income share s_j	Within group $p_j s_j \hat{G}_{Wj}$	Between group $p_j s_j \hat{G}_{Bj}$	Interaction $2p_j s_j \hat{G}_{Ij}$	Total contribution*
Non PR ($j = 1$)	0.6016	−124.2051	96.0433	0.0080	0.0062	0.0000	−0.0061	0.0095	0.0034
Immigrant ($j = 2$)	0.5441	−3.3171	2.5051	0.2279	0.2228	0.0277	−0.1685	0.2545	0.1136
Non-IM. ($j = 3$)	0.5112	0.3088	−0.0681	0.7641	0.7710	0.3011	0.1819	−0.0802	0.4028
Total	–	–	–	1.0000	1.0000	0.3288**	0.0073**	0.1837**	0.5198

*The total contribution of subgroup j to overall inequality $= p_j s_j \hat{G}_{Wj} + p_j s_j \hat{G}_{Bj} + 2p_j s_j \hat{G}_{Ij}, j = 1, 2, 3$. For the nonpermanent resident group ($j = 1$), the total contribution is 0.0034; for the immigrant group ($j = 2$), the total contribution is 0.1136; and for the nonimmigrant group ($j = 3$) the total contribution is 0.4028.

**The total contribution of the within-group component to overall inequality $= \sum_{j=1}^{3} p_j s_j \hat{G}_{Wj} = 0.3288$; the total contribution of the between-group component $= \sum_{j=1}^{3} p_j s_j \hat{G}_{Bj} = 0.0073$; and the total contribution of the interaction component $= 2\sum_{j=1}^{3} p_j s_j \hat{G}_{Ij} = 0.1837$.

breakdown of the overall Gini index into the within-group, between-group and interaction components. Table 6 provides a two-way breakdown of the Gini subgroup decompositions. It is apparent from the entries in the table that the nonimmigrant subgroup makes the highest contribution to the observed inequality even though it has the lowest degree of within-group inequality, with an estimated Gini index of 0.511. Also, the nonpermanent resident subgroup makes the lowest contribution to overall inequality, yet it has the lowest degree of within-group inequality, with an estimated Gini index of 0.602. These results demonstrate one area of similarity between Gini subgroup decomposition, which also provides information on the contributions of the various population subgroups to overall inequality, and Gini income source decomposition, which provides information on the contributions of the various income sources to overall inequality.

The results presented in Table 6 confirm the possibility of negative between-group or interaction pseudo-Ginis as already alluded to above. For example, the table shows that both the immigrant and nonpermanent resident subgroups contribute negatively to between-group inequality, whereas the nonimmigrant subgroup is the sole negative contributor to the interaction component. In this regard, it is interesting to note that the nonpermanent resident and immigrant subgroups that contribute negatively to the between-group component are the ones whose income ranges fall entirely within the income range for the nonimmigrant subgroup, whose contribution to between-group inequality turns out to be positive. Furthermore, the nonimmigrant subgroup for which part of the income range does not overlap with the income ranges of the other two subgroups makes a negative contribution to the interaction component. Clearly, Table 6 illustrates the bidimensional nature of Gini subgroup decomposition, which could be fully exploited in future empirical studies. Specifically, it shows that Gini subgroup decomposition does not only entail a breakdown of overall inequality into within-group, between-group and interaction components but also a breakdown of the contributions of each subgroup to overall inequality, which is akin to the breakdown in income source decompositions.

7. Concluding Remarks

In this article we have extended previous literature in which the regression approach is used to construct the overall Gini index to provide a new way of viewing and decomposing the overall Gini index by population subgroups into within-group, between-group and interaction components. The methodology proposed entails specifying certain underlying "trick regression models" with known heteroscedastic structures related to income.

Although the stochastic approach proposed in this article provides accurate point estimates of the within-group, between-group and interaction components of the Gini decompositions, it would not be adequate to use the associated OLS/WLS standard errors owing to the problems mentioned in Section 5. In light of this difficulty, the use of jackknife or bootstrap standard errors of the subgroup decompositions is recommended.

Finally, we believe that this article demonstrates how the pseudo-Ginis that are computed in the intermediate stages of the subgroup decompositions of the Gini index can provide very useful insights into the contribution of each population subgroup to the within-group, between-group and interaction components of overall inequality and hence the total contribution of each subgroup to overall inequality.

8. References

Bhattacharya, N. and Mahalanobis, B. (1967). Regional Disparities in Household Consumption in India. American Statistical Association Journal, 62, 143–161. DOI: http://www.dx.doi.org/10.1080/01621459.1967.10482896.

Biewen, M. (2002). Bootstrap Inference for Inequality, Mobility and Poverty Measurement. Journal of Econometrics, 108, 317–342. DOI: http://www.dx.doi.org/10.1016/S0304-4076(01)00138-5.

Cowell, F.A. (1989). Sampling Variance and Decomposable Inequality Measures. Journal of Econometrics, 42, 27–41. DOI: http://www.dx.doi.org/10.1016/0304-4076(89)90073-0.

Dagum, C. (1997). A New Approach to the Decomposition of the Gini Income Inequality Ratio. Empirical Economics, 22, 515–531. DOI: http://www.dx.doi.org/10.1007/978-3-642-51073-1_4.

Davidson, R. (2009). Reliable Inference for the Gini Index. Journal of Econometrics, 150, 30–40.

Dixon, P.M., Weiner, J., Mitchel-Olds, T., and Woodley, R. (1987). Bootstrapping the Gini Coefficient of Inequality. Ecology, 68, 1548–1551. DOI: http://www.dx.doi.org/10.2307/1939238.

Fei, J.C.H., Ranis, G., and Kuo, S.W.Y. (1978). Growth and the Family Distribution of Income by Factor Components. Quarterly Journal of Economics, 92, 17–53. DOI: http://www.dx.doi.org/10.2307/1885997.

Giles, D. (2004). Calculating a Standard Error for the Gini Coefficient: Some Further Results. Oxford Bulletin of Economics and Statistics, 66, 425–433. DOI: http://www.dx.doi.org/10.1111/j.1468-0084.2004.00086.x.

Gray, D., Mills, J.A., and Zandvakili, S. (2003). Statistical Inference of Inequality With Decompositions: the Canadian Experience. Empirical Economics, 28, 291–302. DOI: http://www.dx.doi.org/10.1007/s001810200131.

Kanbur, R. (2006). The Policy Significance of Inequality Decompositions. Journal of Economic Inequality, 4, 367–374. DOI: http://www.dx.doi.org/10.1007/s10888-005-9013-5.

Karoly, L.A. (1992). Changes in the Distribution of Individual Earnings in the United States: 1967-1986. Review of Economics and Statistics, 74, 107–115. Available at: http://www.jstor.org/stable/2109548 (access September 1, 2011).

Langel, M. and Tillé, Y. (2013). Variance Estimation of the Gini Index: Revisiting a Result Several Times Published. Journal of the Royal Statistical Society Series A Statistics in Society, 176, 521–540. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2012.01048.x.

Maasoumi, E. (1994). Empirical Analysis of Inequality and Welfare. Handbook of Applied Microeconomics, P. Schmidt and H. Peasaran (eds). Oxford: Blackwell.

Mills, J. and Zandvakili, S. (1997). Statistical Inference Via Bootstrapping for Measures of Economic Inequality. Journal of Applied Econometrics, 12, 133–150. DOI: http://www.dx.doi.org/10.1002/(SICI)1099-1255(199703)12:2<133::AID-JAE433>3.0.CO;2-H.

Modarres, R. and Gastwirth, J.L. (2006). A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality. Oxford Bulletin of Economics and Statistics, 68, 385–390. DOI: http://www.dx.doi.org/10.1111/j.1468-0084.2006.00167.x.

Mussard, S. and Richard, P. (2012). Linking Yitzhaki's and Dagum's Gini Decompositions. Applied Economics, 44, 2997–3010. DOI: http://www.dx.doi.org/10.1080/00036846.2011.568410.

Ogwang, T. (2000). A Convenient Method of Computing the Gini Index and its Standard Error. Oxford Bulletin of Economics and Statistics, 62, 123–129. DOI: http://www.dx.doi.org/10.1111/1468-0084.00164.

Ogwang, T. (2004). Calculating a Standard Error for the Gini Coefficient: Some Further Results: Reply. Oxford Bulletin of Economics and Statistics, 66, 435–437. DOI: http://www.dx.doi.org/10.1111/j.1468-0084.2004.00087.x.

Ogwang, T. (2006). A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality: Comment. Oxford Bulletin of Economics and Statistics, 68, 391–393. DOI: http://www.dx.doi.org/10.1111/j.1468-0084.2006.00167.x.

Ogwang, T. (2007). Additional Properties of a Linear Pen's Parade for Individual Data Using the Stochastic Approach to the Gini Index. Economics Letters, 96, 369–374.

Pyatt, G. (1976). On the Interpretation and Disaggregation of Gini Coefficients. Economic Journal, 86, 243–255. Available at: http://www.jstor.org/stable/2230745.

Radaelli, P. (2010). On the Decomposition by Subgroups of the Gini Index and Zenga's Uniformity and Inequality Indexes. International Statistical Review, 78, 81–101. DOI: http://www.dx.doi.org/10.1111/j.1751-5823.2010.00100.x.

Rao, V.M. (1969). Two Decompositions of Concentration Ratio. Journal of the Royal Statistical Society Series A (General), 132, 418–425. Available at: http://www.jstor.org/stable/2344120.

Sastry, D.V.S. and Kelkar, U.R. (1994). Note on the Decomposition of Gini Inequality. Review of Economics and Statistics, LXXVI, 584–586. Available at: http://www.jstor.org/stable/2109984.

Shao, J. and Tu, D. (1995). Jackknife and Bootstrap. New York: Springer.

Shorrocks, A.F. (1982). Inequality Decomposition by Factor Components. Econometrica, 50, 193–211. Available at: http://www.jstor.org/stable/1912537.

Silber, J. (1989). Factor Components, Population Subgroups and the Computation of the Gini Index of Inequality. Review of Economics and Statistics, 71, 107–115. Available at: http://www.jstor.org/stable/1928057.

Yao, S. and Liu, J. (1996). Decomposition of the Gini Coefficient by Class: A New Approach. Applied Economics Letters, 3, 115–119.

Yao, S. (1999). On the Decomposition of Gini Coefficients by Population Class and Income Source: A Spreadsheet Approach and Application. Applied Economics, 31, 1249–1264. DOI: http://www.dx.doi.org/10.1080/000368499323463.

Yitzhaki, S. (1991). Calculating Jackknife Variance Estimators for Parameters of the Gini Method. Journal of Business and Economic Statistics, 9, 235–239. DOI: http://www.dx.doi.org/10.1080/07350015.1991.10509849.

Received October 2011
Revised May 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 107–122, http://dx.doi.org/10.2478/jos-2014-0006

Disclosure Risk from Factor Scores

Jörg Drechsler[1], *Gerd Ronning*[2], *and Philipp Bleninger*[3]

Remote access can be a powerful tool for providing data access for external researchers. Since the microdata never leave the secure environment of the data-providing agency, alterations of the microdata can be kept to a minimum. Nevertheless, remote access is not free from risk. Many statistical analyses that do not seem to provide disclosive information at first sight can be used by sophisticated intruders to reveal sensitive information. For this reason the list of allowed queries is usually restricted in a remote setting. However, it is not always easy to identify problematic queries. We therefore strongly support the argument that has been made by other authors: that all queries should be monitored carefully and that any microlevel information should always be withheld. As an illustrative example, we use factor score analysis, for which the output of interest – the factor loading of the variables – seems to be unproblematic. However, as we show in the article, the individual factor scores that are usually returned as part of the output can be used to reveal sensitive information. Our empirical evaluations based on a German establishment survey emphasize that this risk is far from a purely theoretical problem.

Key words: Remote data access; confidentiality; statistical disclosure control; factor analysis.

1. Introduction

The scientific community relies heavily on high quality data for the empirical validation of proposed theoretical models. However, data collection is an expensive and laborious task and thus it is prudent to use data which have already been collected by others, albeit for different reasons. Public administrations, governmental agencies and other state institutions gather valuable information on all aspects of society and there are huge benefits to be gained from broad access to these data. The crucial point is how to grant this access without violating the confidentiality guarantees given to survey respondents. Most microdata sets are collected under a pledge of confidentiality and therefore cannot be released unrestrictedly. Statistical analyses via remote access seem to offer both preservation of confidentiality and unlimited use of data. In a remote access system as we define it, the analyst uses his or her desktop computer to connect to a server on which the confidential microdata are stored. He or she can submit any query to the server, which runs the requested analysis of the microdata and returns the results to the user if the requested

[1] Institute for Employment Research, Statistical Methods, Regensburger Str. 104, Nuremberg 90478, Germany. Email: joerg.drechsler@iab.de
[2] University of Tuebingen, Mohlstraße 36, 72074 Tuebingen, Germany. Email: gerd.ronning@uni-tuebingen.de
[3] GfKSE, Nuremberg, Germany.

Acknowledgments: This research was partially supported by the "InfinitE" project funded by the German Federal Ministry of Education and Research. We thank the three referees for their valuable comments, which helped to improve the quality of the article.

output does not violate any confidentiality restrictions. The microdata never leave the secure environment of the server. However, to guarantee that the provided output does not reveal any confidential information, the list of allowed queries is generally limited in practice. The remote access solutions that have been implemented so far either define a list of queries that are not allowed (any command that is not on the list can be requested) or explicitly state which queries can be submitted.

An example of the first approach is the system implemented at the Cross-National Data Center in Luxembourg, known as LISSY (Cross-National Data Center in Luxembourg 2012a), which accepts code written for the software packages SAS, STATA or SPSS. Jobs can be submitted either per e-mail or via a job submission interface. The system does not restrict the list of allowed queries in advance. Instead, "certain syntax and comments will trigger system security alerts" (Cross-National Data Center in Luxembourg 2012b), which may terminate the job. The system will only return ASCII output, that is, no graphical output of any form will be provided. A more advanced version of the approach is also planned in the U.S. (Lucero et al. 2011).

An example of the second approach is implemented at the National Center for Health Statistics (NCHS) (Research Data Center of the National Center for Health Statistics 2012a). The NCHS system, which is called ANDRE, only accepts code written for the software packages SAS or SUDAAN. Other software packages, such as SPSS or R, can only be used on-site. Furthermore, the list of possible procedures and options is limited in advance and some procedures will automatically be adapted to avoid disclosure (Research Data Center of the National Center for Health Statistics 2012b). Finally, the website states that "[o]utput results that pose a disclosure risk will be suppressed" without any further information as to how such an output is identified. This kind of approach has also been implemented in Australia (O'Keefe and Good 2008). The Australian Bureau of Statistics provides an online tool called TableBuilder "which enables users to create tables, graphs and maps of Census data" (http://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder). Another online tool called DataAnalyser which additionally allows running a number of standard regression models will also be implemented soon (http://www.abs.gov.au/websitedbs/D3310114.nsf/home/About+DataAnalyser).

However, even though the list of allowed queries is generally limited in a remote access setting to avoid disclosure from simple attacks like maximum queries, some attacks are harder to detect, especially if these attacks are based on multivariate analysis. One of the more prominent examples is the disclosure risk from linear regression. Gomatam et al. (2005) describe two possible strategies that an intruder with background knowledge about some of the survey respondents can apply to obtain any sensitive information contained in the data set regarding these survey respondents. Bleninger et al. (2011) further formalize these strategies and apply them to a German establishment survey. They find that very limited background information is sufficient to obtain exact information on sensitive attributes in the data set. Since the risks from linear regression are well known in the SDC community, the current implementations of remote access already take measures to ensure that these strategies cannot be applied. However, this highlights the essential dilemma of the remote access environment: Possible intruder strategies need to be known in advance to enable the implementation of counterstrategies. Restricted remote access following the second approach described above is an attempt to circumvent this dilemma by only

allowing computations that are considered safe under all circumstances. However, as a consequence, the set of allowed queries will be very limited and many users will find this set too restricted to answer their respective research question. Thus, for most researchers full remote access is the only viable solution. In this context, full remote access would mean that only those queries that are known to be disclosive would be prohibited. However, implementing such a fully automated approach would mean that all potentially risky queries are known in advance so that the number of suppressed queries can be kept to a minimum. This is an ambitious goal and it is not clear whether this goal can ever be achieved.

While the risks from releasing microlevel information of the original data are obvious, it is less obvious that microlevel information is a byproduct of several data analysis tools and that this byproduct might pose a risk although the final output of interest might not be problematic. Regression procedures provide microlevel output such as fitted values or residuals, and model-fitting checks, such as Q-Q plots or Cook's distance, provide information on the individual level at least for the outliers (arguably the most interesting individuals for an intruder). Although at first sight it seems impossible to learn anything about the reported microdata values from these diagnostic plots, Sparks et al. (2008) illustrate the risks that might result if these analytics tools are provided in a remote access system without further restrictions. For this reason, the remote access system that is planned for the U.S. will, for example, provide Q-Q plots that are based on synthetic data. Sparks et al. (2008) also suggest a number of additional protective measures that can be taken to avoid these kinds of disclosures and argue that no information on the individual level should be released in general. To our knowledge, all agencies that have implemented a remote access environment so far have followed this advice.

In this article we provide another example of why monitoring the output of any analysis and suppressing all microlevel information is generally a good strategy. Factor analysis is very popular in the social sciences since it can be applied in a wide range of explorative and confirmatory tasks and it would be a severe drawback of remote access if this kind of analysis was not possible. On the other hand, as we will illustrate in this article, there is a risk of disclosure if unrestricted factor analysis is allowed. However, this risk can easily be avoided if the individual factor scores are not revealed to the analyst. Since researchers will usually only be interested in the factor loadings for the different variables included in the model, we do not see any disadvantages in not providing the individual factor scores. If information on the individual factors is considered necessary, graphical displays of the winsorised data could be provided akin to the disclosure prevention measures described in Sparks et al. (2008).

The remainder of the article is organized as follows: Following a brief description of factor analysis methods, we provide a short overview of different estimation procedures for factor scores. Section 4 demonstrates that there is a risk of disclosure for all these approaches if a set of variables could be identified in the data set that is uncorrelated with the variable to be disclosed, henceforth called the variable of interest. The empirical example in Section 5 shows that such a correlation structure is not uncommon in practice and once the "appropriate" set of variables is selected, it is possible to estimate the true values for every record in the data set very precisely for the variable of interest. The data

for this empirical illustration are taken from the IAB Establishment Panel, a survey conducted by the Institute for Employment Research (IAB) in Germany. The article concludes with some final remarks.

2. Some Basic Facts on Factor Analysis

Factor analysis and the closely related method of principal components are widely used in all fields of social science, in particular in psychology and sociology where "latent" variables, such as ability and satisfaction, are modelled frequently. More recently, the method has also been employed in modern time series analysis when factor-augmented vector autoregression models (FAVAR) are considered (see, for example, Stock and Watson 2002). The aim of the approach is to reduce the empirical information from a large set of continuous variables to a small set of (latent) factors. In the following we describe the basic concept briefly. A detailed description can be found in any standard textbook on the topic (see, for example, Press 2005).

Consider a set of m random variables $\eta = (\eta_1, \eta_2 \ldots, \eta_m)'$ with

$$E[\eta] = \mu_\eta, cov[\eta] = \Sigma_{\eta\eta}$$

for which n observations are available leading to the $(n \times m)$ data matrix $Y = (y_1, y_2, \ldots, y_m)$. The factor model seeks to explain the m variables by a set of $p < m$ "common factors" $\mathbf{f} = (f_1, f_2, \ldots, f_p)'$ through the linear model

$$\eta - \mu_\eta = \Lambda \mathbf{f} + \mathbf{u}, \tag{1}$$

where Λ is the $(m \times p)$ factor-loading matrix and $\mathbf{u}$ is an m-dimensional vector of "specific factors" with

$$E[\mathbf{u}] = 0, cov[\mathbf{u}] = \Psi = \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_m \end{pmatrix}.$$

Since the factors are assumed to be orthogonal with $cov[\mathbf{f}] = I_p$, where I_p is the identity matrix of dimension $(p \times p)$, as well as independent of $\mathbf{u}$, we obtain what is called the "fundamental equation"

$$\Sigma_{\eta\eta} = \Lambda\Lambda' + \Psi.$$

Let F be the $(n \times p)$-matrix of realized factor scores which is related to the data matrix Y by the equation (McDonald and Burr 1967, p. 384)

$$Y - M = F\Lambda' + U, \tag{2}$$

which implicitly defines the $(n \times m)$ matrix M by

$$M = \iota_n \otimes \mu_\eta'.$$

Here ι_n is an n-vector of ones and $\otimes$ denotes the Kronecker product, that is, the corresponding mean from the vector μ_η is subtracted from each observation in Y in (2).

We will call (2) the "empirical factor model", whereas (1) will be called the "theoretical factor model".

If the estimated matrix Λ has a block-diagonal structure, particular factors can be related to a subset of the vector η, which helps to interpret these factors. However, it is well known that this estimated matrix is not unique: Take any $(m \times m)$ orthogonal matrix W and it will by definition satisfy $WW' = I_m$. Keeping this in mind, we can rewrite (1) as

$$\eta - \mu_\eta = (\Lambda W)(W'\mathbf{f}) + \mathbf{u},$$

where $\Lambda^* = \Lambda W$ would represent the factor-loading matrix and $\mathbf{f}^* = W'\mathbf{f}$ the vector of factors. The multiplication of the factor-loading matrix by any orthogonal matrix is called rotation of this matrix. Usually, the matrix W is chosen such that for each factor the loading on a subset of variables is as large as possible and the loading on the remaining variables is as small as possible, so that a "simple structure" is obtained which facilitates the interpretation of factors. One way to achieve this is to find the orthogonal matrix that maximizes the variance of the squared factor loadings. This is the well-known varimax criterion (see, for example, Press 2005 Ch. 10.6 for details).

3. Estimation of Factor Scores

This section provides a short review of the four different approaches that are discussed in the literature for obtaining factor scores (see Ronning and Bleninger 2011 for a more detailed review that also presents the derivations for all estimators). In the following we assume that the factor-loading matrix Λ is known or rather has been estimated in an earlier step indicated by the symbol $\sim$ placed above the relevant quantities. Hence, the resulting estimates of $\mathbf{f}$ depend on the method by which the factor-loading matrix was determined. In all cases $\tilde{\Lambda}$ may represent either the original or the rotated factor-loadings. We will only present the results for the empirical model (2) as this will be the relevant model for our disclosure risk evaluations in the following sections. Derivation of the results for the theoretical model (1) is straightforward.

3.1. Least Squares Solution

The empirical factor model (2) can be seen as a regression model with unknown matrix F which can be estimated by least squares. The resulting estimator is

$$\hat{F}_{LS} = (Y - M)\tilde{\Lambda}(\tilde{\Lambda}'\tilde{\Lambda})^{-1}. \tag{3}$$

Note that the transpose of $\hat{F}_{LS}$ is just the standard OLS estimate from linear regression. Horst (1965) seems to have been one of the first to use this approach (McDonald and Burr 1967, p. 386).

3.2. Bartlett's Method

Considering the nonscalar structure of the covariance matrix Ψ, a generalized least squares formula seems more appropriate:

$$\hat{F}_{BA} = (Y - M)\tilde{\Psi}^{-1}\tilde{\Lambda}(\tilde{\Lambda}'\tilde{\Psi}^{-1}\tilde{\Lambda})^{-1}. \tag{4}$$

Note that in this case the matrix Ψ also has to be determined in advance. This method has been proposed by Bartlett (1937). Fahrmeir et al. (1996, pp. 648, 690) remark that (4) can be regarded as a maximum likelihood estimator when normality for η is assumed. Non-normally distributed variables in η lead to quasi-maximum likelihood estimation of loadings and scores, still being asymptotically normally distributed and consistent.

3.3. Thomson's Method

The method is attributed to both Thomson (1939) and Thurstone (1935). Thurstone (1935) derived the factor scores by requiring that the estimated factor score $\hat{f}_j$ be as close to the "true" factor score f_j as possible for $j = 1, \ldots, p$. He considers the linear estimator

$$\hat{f}_j = \mathbf{a}_j'(\eta - \mu)$$

for which the mean-squared error should be minimized with respect to the vector $\mathbf{a}_j$ (see Ronning and Bleninger 2011 for details). With this approach, the factor scores in the empirical model are given by:

$$\hat{F}_{TH} = (Y - M)\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right)^{-1}\hat{\Lambda}. \tag{5}$$

3.4. Principal Component Analysis

Of course, the principal component approach can also be used to estimate the factor scores: If we consider the spectral decomposition of the covariance matrix

$$\Sigma_{\eta\eta} = Q\Theta Q',$$

the principal components $\boldsymbol{p}_j,\ j = 1, \ldots, m$, are given by the matrix

$$\left(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{m-1}, \mathbf{p}_m\right) = P = YQ = \left(Y\mathbf{q}_1, Y\mathbf{q}_2, \ldots, Y\mathbf{q}_{m-1}, Y\mathbf{q}_m\right),$$

where the columns $\mathbf{q}_j$ are the characteristic vectors of the covariance matrix, whereas the diagonal matrix Θ contains the characteristic values. Usually, only the principal components corresponding to the largest characteristic values are used since they represent maximum variation. The matrix P can be seen as the matrix of estimated factors, that is,

$$\hat{F}_{PC} = P. \tag{6}$$

For more details see any textbook on multivariate analysis, such as, Press (2005).

4. Disclosure Risk from Factor Analysis

In this section we will illustrate scenarios in which the factor scores disclose sensitive information. We show analytically that a severe risk of disclosure exists if at least one variable can be identified in the data set that is (almost) uncorrelated with the variable of interest. As we show later in the empirical example (Subsection 5.3), potential variables can be selected by inspecting the correlation matrix.

For concreteness, let us assume that η_1 is the variable of interest so that the covariance matrix has the following block diagonal structure:

$$\Sigma_{\eta\eta} = \begin{pmatrix} \sigma_{11} & 0' \\ 0 & \Sigma_{22} \end{pmatrix} \tag{7}$$

where Σ_{22} is the $(m-1)\times(m-1)$ covariance matrix of the remaining $m-1$ variables. Clearly, this leads to a factor-loading matrix with one factor "loading" only on the first variable and the remaining $p-1$ factors having zero loading weight on this variable. Note that this implies

$$(\Lambda'\Lambda)^{-1} = \begin{pmatrix} 1 & 0' \\ 0 & (\Lambda_2'\Lambda_2)^{-1} \end{pmatrix} \tag{8}$$

where Λ_2 is the $m\times(p-1)$ loading matrix of the remaining $p-1$ variables.

Substituting (8) into (3) for the least squares solution and into (4) for Bartlett's method, we obtain identical results regarding the uncorrelated variable (the derivations are presented in the Appendix)

$$F_{LS} = F_{BA} = \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right).$$

Therefore, for both the least squares solution and Bartlett's method, the first factor $\mathbf{f}_1$ is identical (up to an additive constant) to the data vector $\mathbf{y}_1$ and it will be easy for the intruder to derive the values for $\mathbf{y}_1$ at least approximately, since computing the mean of a variable will usually be allowed in a remote access environment. Note that only the first factor $\mathbf{f}_1$ is identical for the least squares solution and for Bartlett's method. The estimated factors for $j = 2, \ldots, p$ will generally differ for the two methods. For the solution of Thomson/Thurstone we obtain a slightly different result (again, derivations are presented in the Appendix):

$$F_{TH} = \left(\frac{1}{1+\psi_1} \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right).$$

The results show that in this case the estimated factor $\mathbf{f}_1$ not only differs by an additive constant, but the multiplicative factor $1/(1+\psi_1)$ also has to be taken into account. If ψ is small or the estimate of ψ used in the computation is available, disclosure risk is high.

Finally, for the principal component approach one of the characteristic values, say θ_j, equals σ_{11}. The corresponding characteristic vector must then satisfy $\mathbf{q}_j = (1\ 0.\ .\ .0)'$. Therefore, the corresponding principal component is given by

$$\mathbf{p}_j = {}^{Y}\mathbf{q}_j = \mathbf{y}_1,$$

so that in this case the data vector $\mathbf{y}_1$ is exactly reproduced by the principal component. It should be noted, however, that θ_j is not necessarily the largest characteristic value (see Ronning and Bleninger 2011 for a formal proof). Since usually only the principal components corresponding to the largest characteristic values are used in practice, extracting the vector for components corresponding to small characteristic values might be suspicious and agencies might prevent some attacks based on this approach if only the components corresponding to the largest characteristic values can be retrieved.

As a final remark, we try to shed some light on the question of what influence the $m-1$ "remaining" variables in the factor model have on the accuracy of the results. Most importantly, whenever at least one variable highly correlated with the variable of interest is included in the factor model, there will be no eigenvector loading on the variable of interest alone and no disclosure will be possible. Clearly, if the correlation with the variable of interest is exactly zero for all variables included in the set of variables in m, the theoretical results above imply exact reproduction of the vector $\mathbf{y}_1$ no matter how many additional variables are included in the set of variables in m. In this case, one variable would be sufficient and adding variables that are (even slightly) correlated with the variable of interest will decrease the level of accuracy. In practice, the correlation is never exactly zero, as illustrated in Table 1 from our empirical example in Section 5. However, it would still make sense in terms of prediction accuracy to only pick the variable with the lowest correlation with the variable of interest. Nevertheless, it might generally be advantageous from the perspective of an intruder to include some additional variables in the model to avoid submitting queries that look overly suspicious. In this case it would be the best strategy to pick a predefined set of variables, say eight to ten, consisting of those variables with the lowest empirical correlation with the variable of interest. This is the strategy we follow in our empirical evaluations in the next section.

5. Empirical Evidence

5.1. The Data

The IAB Establishment Panel is a nationwide annual survey of establishments in Germany conducted by the Institute for Employment Research (IAB). It includes establishments with at least one employee covered by social security and contains business-related facts (e.g., management, business policy, innovations), a large number

of employment policy-related subjects (e.g., personnel structure, recruitment, wages and salaries) and a range of background information (e.g., regional information, industrial sector). For further information see Fischer et al. (2009) and Kölling (2000). The IAB collects the data under the pledge of confidentiality. Additionally, German law restricts the release of data from public administrations to avoid the disclosure of sensitive information. Therefore, direct access to the survey is only granted to external researchers at the IAB's research data center (RDC). The RDC, which was established in 2004, provides researchers with access to microdata for noncommercial empirical research in the fields of social security and employment. Most of the surveys conducted at the IAB and samples from the administrative data of the Federal Employment Agency are available for on-site analysis (see Heining 2009 or http://fdz.iab.de for further details).

Researchers can also submit queries to the RDC that are run on the original data by the staff of the RDC (remote execution). In this case the results are reported back to the researchers only after the output has been carefully checked for confidentiality violations (if the researcher analyzes the data onsite, only the results that are intended to be used outside of the rooms of the RDC will be checked). Finally, some surveys are also available as scientific use files (unlike public use files, scientific use files are only available to the scientific community). Currently, all confidentiality checks are performed manually, so the attack described in this article would be detected. Nevertheless, as remote access is seen to be the future for data providers, we use the data set to illustrate that unrestricted factor analysis in a remote access setting would be problematic in terms of disclosure risk.

For our empirical evaluations we use the cross-section from the year 2007 of the survey. All missing values in this data set are replaced by single imputation and treated as observed values. See Drechsler (2011) for a description of the imputation of the missing values in the survey. The sensitive variable to be disclosed is the turnover from an establishment's sales after taxes, that is, the revenue. Thus, we exclude all establishments that do not report turnover, such as nonindustrial organizations, regional and local authorities and administrations, financial institutions and insurance companies. The remaining data set includes 12,814 fully observed establishments.

5.2. *Estimation of Factor Loadings*

Since the very skewed distribution of the turnover variable generates some outliers among the factor scores, we transform the variable according to

$$lgturn_i = \log(turnover_i + 1), \tag{9}$$

where $turnover_i$ is the turnover in euros for establishment i. The 1 is added to ensure that all values are strictly positive before the log transformation, because some establishments report a turnover of zero. The transformed variable is approximately normally distributed, leading to approximately unbiased and consistent maximum likelihood estimation of the corresponding loadings and scores for Bartlett's method.

In order to successfully apply the disclosure attack outlined above, we need to identify variables that are (almost) uncorrelated with this variable. It should not be difficult for an intruder to obtain this information because correlation matrices are not usually considered

Table 1. Variables used in the factor scores model

Variable	$\rho(lgturn, y_j)$
Turnover from sales after taxes on the log scale (lgturn.)	1.0000
Investments in IT (inv.)	0.0587
Total number of civil servant aspirants (asp.)	0.0082
Total number of vacant positions for workers (vac.w.1)	0.0536
Number of vacancies for workers reported to employment agency (vac.w.2)	0.0374
Number of vacancies for qualified employees reported to employment agency (vac.em.)	0.1193
Employees with wage subsidies (sub.)	0.0984
Employees over 50 with wage subsidies (sub.50)	0.0513

to provide a high risk of disclosure. Table 1 lists the eight variables that we use in the factor scores model together with their empirical correlation with the log turnover ($\rho(lgturn)$).

Of course the assumption of zero correlation underlying the results in Section 4 is unrealistic for real data settings, but the correlations in Table 1 are small and we will see that the originally reported turnover can still be estimated almost exactly with this scenario.

Usually, factor analysis starts by inspecting the eigenvalues of the covariance matrix or correlation matrix to determine the number of factors p to be used in the model. Only the largest eigenvalues are selected with the understanding that the variability of Y is sufficiently explained by this subset. Based on the correlation matrix both the Kaiser criterion (Kaiser 1958) and the scree test (Fahrmeir et al. 1996) would suggest selecting $p = 4$ for our set of variables. However, inspecting the eigenvalues is not helpful in our setting since we need to make sure that the factor that loads on the variable of interest alone is also included in the model. As noted earlier, it can be shown that the relevant eigenvalue need not be one of the largest eigenvalues (see Ronning and Bleninger 2011 for more details). Therefore, the intruder should choose a large $p \leq m$ and examine all estimated factors. Alternatively, he or she could simply try alternative values of p. We found the ideal number of factors by evaluating the full range of possible factors. The loading matrix for $p = 4$ (after rotation based on the varimax criterion) is presented in Table 2 and it is obvious that in this case the third factor loads primarily on turnover and thus this factor model is ideally suited for a disclosure attack.

Table 2. Rotated Matrix $\tilde{\Lambda}$ of estimated loadings

	Factor 1	Factor 2	Factor 3	Factor 4
lgturn.	0.0202	0.0360	0.9867	0.1406
inv.	−0.0046	0.0019	0.0326	0.1888
asp.	0.0002	0.0051	0.0105	−0.0167
vac.w.1	0.9879	0.0134	0.0267	0.0487
vac.w.2	0.9325	0.0090	0.0089	0.0673
vac.em.	0.0796	0.0742	0.0853	0.2194
sub.	0.0166	0.7933	0.0719	−0.0100
sub.50	0.0041	0.9958	0.0088	0.0471

5.3. Estimation of Factor Scores

In the next step, we estimate the matrix of factor scores $\hat{F}$ based on the rotated loadings from Table 2. For purpose of brevity, we limit our evaluation to Bartlett's (4) and Thomson's (5) solution. We note that the least squares solution and principal component analysis will provide similar results. Once we have estimated the score values, we obtain the estimated values for the transformed turnover variable by adding its mean to all the factor scores based on the assumption that the mean of the (transformed) variable is available in remote access. To approximate turnover on the original scale, we transform the obtained values according to

$$\hat{turn}_i = exp\left\{\widehat{lgturn}_i\right\} - 1.$$

We note that the transformation will lead to a small bias in the estimated turnover since in general $E(\log(y_i)) \neq \log(E(y_i))$. To evaluate how close the resulting estimate is to the reported turnover, we use the difference between reported and estimated turnover relative to the reported turnover

$$\delta_i = \frac{\hat{turn}_i - turnover_i}{turnover_i}, \quad i = 1, \ldots, n.$$

The two leftmost panels in Figure 1 show scatter plots of these differences for Bartlett's (left panel) and Thomson's method (middle panel) respectively. In the scatter plots, the establishments are sorted in ascending order based on the number of employees.

Looking at the scatter plots, we find that using Bartlett's method the estimated turnover is very close to the true turnover for almost all establishments. The relative difference δ is less than 0.5% for 99.3% of the establishments.

For Thomson's method, we notice that the relative differences are generally larger than for Bartlett's method (note that the scale of δ differs between the scatter plot for Thomson's method (middle panel) and the scatter plots for Barlett's method (left panel) and Thomson's method after correction (right panel)). More than 40% of the estimated

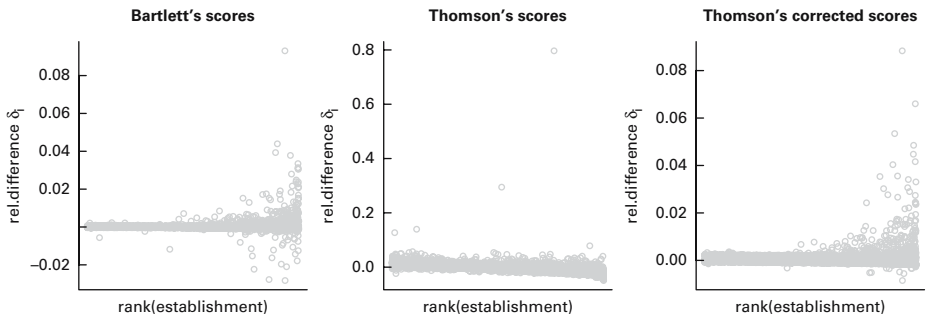

Fig. 1. Relative differences δ_i between the true turnover and the turnover estimated from the factor scores obtained through Bartlett's method and Thomson's method with/without correction. Establishments are sorted in ascending order based on the number of employees.

turnovers differ by more than 1% from the true value and the difference can be up to 80%. We also find a trend in the relative differences. The turnover derived from the factor scores overestimates the true turnover for the smallest establishments. This effect is decreases continuously and the turnover is underestimated for large establishments. This is not surprising if we note that we obtained our estimate for turnover by adding the sample mean to the factor scores without correcting for the multiplicative factor $1/(1+\psi_1)$. Thus, assuming that y_1 is the transformed turnover according to (9) and $\hat{y}_1$ is its estimate based on Thomson's method without correction, the difference between the two quantities is given by

$$\hat{y}_1 - y_1 = \frac{\psi_1}{1+\psi_1} \begin{pmatrix} \mu_1 - y_{11} \\ \vdots \\ \mu_1 - y_{1n} \end{pmatrix}, \tag{10}$$

which will be positive for all establishments with a turnover that is smaller than the average turnover and negative for the rest. Since turnover is highly correlated with establishment size, we observe a negative trend for the relative difference when going from the smallest establishments to the largest. If an estimate for the specific factor $\tilde{\psi}_1$ is available, we can correct the estimator for the reported turnover. The right panel in Figure 1 presents the results based on the corrected estimate. The relative difference δ again is close to zero for almost all establishments, with 99.0% of the establishments, having a relative difference of less than $\pm 0.5\%$. In fact, the estimated turnover never differs by more than $\pm 8.9\%$ from the true turnover. Thus the risk of disclosure is comparable to the risk when Bartlett's method is applied.

6. Conclusions

There is an increasing demand among researchers for access to microdata that have been collected under the pledge of confidentiality. One promising approach to granting access without violating confidentiality guarantees is remote access. However, even though the researcher never has direct access to the underlying microdata, the approach is not free from the risk of disclosure. In our article we have illustrated this risk for a specific analysis that is commonly used in the social sciences: factor analysis. Even though factor analysis is used for information reduction and the potential risk of disclosure is anything but obvious, we showed analytically that individual microlevel values could be obtained exactly for any variable for which a set of covariables can be identified that are uncorrelated with the target variable. This result holds irrespective of the method used to estimate the factor scores. Of course, zero correlation is unrealistic in practice but our empirical example illustrates that a very close approximation of the microlevel values could be obtained even if a small correlation exists between the target variable and the other variables used in the factor model.

It is important to note at this point that by applying the procedure outlined in this article, the intruder will only obtain a full vector of estimated microlevel values. Even if these estimates are very close to the true values, this will not necessarily lead to disclosure if the intruder is not able to link this information to individual units in other databases.

Still, most legislation requires that no individual information be released to the public, no matter whether a direct link is possible or not. Furthermore, it is often easy to attribute some of the obtained values to specific units, such as the largest turnover in the data set, for example.

Finally, we wish to stress that it is not the aim of this article to call for more restrictive data access. Factor analysis is a useful and widely used method that should be available to researchers in a remote access system. We only wish to raise awareness of the fact that this kind of attack is possible if no countermeasures are taken. Once identified, these attacks can be prevented easily by not reporting individual factor scores, since applied analysts are not usually interested in these scores. Following Brandt et al. (2010), who provided general guidelines for output checking when data are disseminated, the factor loadings of the different variables can be considered "safe" outputs that can be released without restrictions. The individual factor scores, on the other hand, should be classified as "unsafe", and extra measures are necessary if these scores are to be provided. Simply checking the correlation between the factor scores and the variables in the data set, for example, could be a useful tool for avoiding disclosure. The factor scores can be suppressed if the bivariate correlation with any variable in the data set is higher than an agency-defined threshold, say 0.995. Alternatively, preventive measures, such as providing only graphical displays of the winsorised factor scores or other measures akin to the measures suggested by Sparks et al. (2008), could be implemented. Finally, as suggested by one of the referees, output perturbation could also be applied. As the name indicates, this approach guarantees confidentiality by only perturbing the output of the queries; the underlying microdata are not altered. This approach has been discussed for other query types such as survival analysis (see, for example O'Keefe et al. 2012) and the original setup for ε-differential privacy (Dwork 2006) was also developed around this idea. Identifying the best perturbation approach when providing individual factor scores would be an area for future research. The aim of this article was more generally to illustrate that data providers granting access to sensitive data should be aware that there are many ways to obtain sensitive information without direct access to the microdata using standard analyses, and not all of them are obvious.

Appendix. Derivations of the Factor Scores if One Variable is Uncorrelated With the Other Variables in the Model

The Least Squares Solution

$$F_{LS} = (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & (\Lambda_2'\Lambda_2)^{-1} \end{pmatrix} = (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2(\Lambda_2'\Lambda_2)^{-1} \end{pmatrix}$$

$$= \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

Bartlett's Method

$$\hat{F}_{BA} = (Y - M)\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}$$

$$= (Y - M)\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\left(\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}'\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\right)^{-1}$$

$$= (Y - M)\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Psi_2^{-1}\Lambda_2 \end{pmatrix}\left(\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Lambda_2'\Psi_2^{-1}\Lambda_2 \end{pmatrix}\right)^{-1}$$

$$= (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Psi_2^{-1}\Lambda_2\left(\Lambda_2'\Psi_2^{-1}\Lambda_2\right)^{-1} \end{pmatrix}$$

$$= \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

The Solution of Thomson/Thurstone

$$F_{TH} = (Y - M)\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right)^{-1}\hat{\Lambda}$$

$$= (Y - M)\left(\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2\Lambda_2' \end{pmatrix} + \begin{pmatrix} \psi_1 & \\ & \Psi_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}$$

$$= (Y - M)\begin{pmatrix} (1 + \psi_1)^{-1} & 0' \\ 0 & \left(\Lambda_2\Lambda_2' + \Psi_2\right)^{-1}\Lambda_2 \end{pmatrix}$$

$$= \left(\frac{1}{1 + \psi_1} \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

7. References

Bartlett, M. (1937). The Statistical Conception of Mental Factors. British Journal of Psychology, 28, 97–104. DOI: http://www.dx.doi.org/10.1111/j.2044-8295.1937.tb00863.x

Bleninger, P., Drechsler, J., and Ronning, G. (2011). Remote Data Access and the Risk of Disclosure from Linear Regression. SORT, Special Issue: Privacy in Statistical Databases, 7–24.

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M., Mol, J., Ritchie, F., Seri, G., and Welpton, R. (2010). Guidelines for the Checking of Output Based on Microdata Research. Final report of ESSnet sub-group on output SDC.

Cross-National Data Center in Luxembourg (2012a). Available at: http://www.lisdatacenter.org (accessed January 17, 2014).

Cross-National Data Center in Luxembourg (2012b). Available at: http://www.lisdatacenter.org/data-access/lissy/best-practices/ (accessed January 17, 2014).

Drechsler, J. (2011). Multiple Imputation in Practice – a Case Study Using a Complex German Establishment Survey. Advances in Statistical Analysis, 95, 1–26. DOI: http://www.dx.doi.org/10.1007/s10182-010-0136-z

Dwork, C. (2006). Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming (ICALP), 1–12.

Fahrmeir, L., Hamerle, A., and Tutz, G. (1996). Multivariate Statistische Verfahren, (2nd edn). Berlin: De Gruyter.

Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2009). The IAB Establishment Panel – Things Users Should Know. Schmollers Jahrbuch – Journal of Applied Social Science Studies, 129, 133–148. DOI: http://www.dx.doi.org/10.3790/schm.129.1.133

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. Statistical Science, 20, 163–177. DOI: http://www.dx.doi.org/10.1214/088342305000000043

Heining, J. (2009). The Research Data Centre of the German Federal Employment Agency: Data Supply and Demand Between 2004 and 2009. RatSWD working paper, 129.

Horst, P. (1965). Factor Analysis of Data Matrices. New York: Holt, Rinehart & Winston.

Kaiser, H. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika, 23, 3, 187–200. DOI: http://www.dx.doi.org/10.1007/BF02289233

Kölling, A. (2000). The IAB-Establishment Panel. Journal of Applied Social Science Studies, 120, 291–300.

Lucero, J., Freiman, M., Singh, L., You, J., DePersio, M., and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau. SORT, Special Issue: Privacy in Statistical Databases, 77–98.

McDonald, R. and Burr, E. (1967). A Comparison of Four Methods for Constructing Factor Scores. Psychometrika, 32, 381–401. DOI: http://www.dx.doi.org/10.1007/BF02289653

O'Keefe, C., Sparks, R., McAullay, D., and Loong, B. (2012). Confidentialising Survival Analysis Output in a Remote Data Access System. Journal of Privacy and Confidentiality 4. Available at: http://repository.cmu.edu/jpc/vol4/iss1/6 (accessed January 17, 2014).

O'Keefe, C.M. and Good, N.M. (2008). A Remote Analysis Server – What Does Regression Output Look Like? In Privacy in Statistical Databases, J. Domingo-Ferrer and Y. Saygin (eds), vol 5262 of Lecture Notes in Computer Science. New York: Springer, 270–283.

Press, S. (2005). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, (2nd edn). New York: Dover Publications.

Research Data Center of the National Center for Health Statistics (2012a). Available at: http://www.cdc.gov/rdc/B2AccessMod/ACs230.htm (accessed January 17, 2014).

Research Data Center of the National Center for Health Statistics (2012b). Available at: http://www.cdc.gov/rdc/Data/B2/SASSUDAANRestrictions.pdf (accessed January 17, 2014).

Ronning, G. and Bleninger, P. (2011). Disclosure Risk From Factor Scores. Technical Report, IAW Discussion Papers 73. Available at: http://www.iaw.edu/w/IAWPDF.php?id=886&name=iaw_dp_73.pdf (accessed January 17, 2014).

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T., and McAullay, D. (2008). Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-preserving Analytics. Comput Methods Programs Biomed, 91, 208–222. DOI: http://www.dx.doi.org/10.1016/j.cmpb.2008.04.001

Stock, J. and Watson, M. (2002). Forecasting Using Principal Components From a Large Number of Predictors. Journal of the American Statistical Association, 97, 1167–1179. DOI: http://www.dx.doi.org/10.1198/016214502388618960

Thomson, G. (1939). The Factorial Analysis of Human Ability. London: University of London Press.

Thurstone, L. (1935). The Vectors of Mind. Chicago: University of Chicago Press.

Received May 2012
Revised April 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 123–146, http://dx.doi.org/10.2478/jos-2014-0007

Disclosure-Protected Inference with Linked Microdata Using a Remote Analysis Server

James O. Chipperfield[1]

Large amounts of microdata are collected by data custodians in the form of censuses and administrative records. Often, data custodians will collect different information on the same individual. Many important questions can be answered by linking microdata collected by different data custodians. For this reason, there is very strong demand from analysts, within government, business, and universities, for linked microdata. However, many data custodians are legally obliged to ensure the risk of disclosing information about a person or organisation is acceptably low. Different authors have considered the problem of how to facilitate reliable statistical inference from analysis of linked microdata while ensuring that the risk of disclosure is acceptably low. This article considers the problem from the perspective of an Integrating Authority that, by definition, is trusted to link the microdata and to facilitate analysts' access to the linked microdata via a remote server, which allows analysts to fit models and view the statistical output without being able to observe the underlying linked microdata. One disclosure risk that must be managed by an Integrating Authority is that one data custodian may use the microdata it supplied to the Integrating Authority and statistical output released from the remote server to disclose information about a person or organisation that was supplied by the other data custodian. This article considers analysis of only binary variables. The utility and disclosure risk of the proposed method are investigated both in a simulation and using a real example. This article shows that some popular protections against disclosure (dropping records, rounding regression coefficients or imposing restrictions on model selection) can be ineffective in the above setting.

Key words: Confidentiality; remote analysis; record linkage; statistical disclosure control.

1. Introduction

Large amounts of microdata are collected by data custodians in the form of censuses and administrative sources. Often, data custodians will collect different information on the same individual. Many important questions can be answered by linking microdata collected by different data custodians. For this reason, there is very strong demand from analysts, within government, business and universities, for linked microdata. However, many data custodians are legally obliged to ensure the risk of disclosing information about a person or organisation is acceptably low. For simplicity, in the rest of this article it is assumed that there are only two data custodians and the *linked microdata* are the result of linking two sets of microdata collected by the two data custodians. Potential analysts of

[1] Senior Research Fellow, National Institute for Applied Statistics Research Australia, University of Wollongong, and Assistant Director, Methodology Division, Australian Bureau of Statistics, Canberra, ACT, 2617, Australia. Email: james.chipperfield@abs.gov.au

the linked microdata are the two data custodians and noncustodians (e.g., academics, members of the public). There are two reasons the disclosure risks are significantly greater if an analyst of the linked microdata is also a data custodian. First, because data custodians commonly collect name and address, any additional information that can be inferred about a record on the microdata it collected, can be directly associated with the person who provided it. Second, a data custodian can use information on the linked microdata collected to disclose information about a person or organisation on the linked microdata that was collected by the other data custodian.

There has been some work in the literature on managing the disclosure risks from analysts who are also data custodians. When unique identifiers, such as name and address, are available, record linkage techniques (see Herzog et al. 2007) are frequently used to identify records belonging to the same individual. Secure Record Linkage (SRL) (see, for example, Churches and Christen 2004) suggests a way in which a third party can link microdata without each data custodian disclosing the identity of nonlinked records to the other data custodian and without the data custodians revealing any sensitive information to the third party. The data custodians attach a unique record identifier to their microdata (e.g., random number) and agree on a common way of encrypting the linking variables, which are sent to the third party to perform the record linkage. The third party links the microdata and returns the record identifiers of the linked pairs to the data custodians. Therefore, each data custodian could identify the names and addresses of the people who were linked, which in turn could disclose sensitive information (e.g., knowing a person's record has been linked to an unemployment register discloses the person is unemployed). For many data custodians, such as the Australian Bureau of Statistics (ABS), revealing such information would be a breach of their legal obligations and would mean that SRL is not a viable option. If instead the third party was allowed access to linking variables (e.g., name and address), the linkage could be of much higher quality, since clearly unencrypted linking variables are more useful in identifying matches than encrypted linking variables. It would be interesting to study the extent to which encryption of linking variables reduces the quality of the linkage.

Once the linked pairs are determined, each data custodian will need to ensure that any statistical output from the linked microdata has an acceptable disclosure risk. Secure computation algorithms allow data custodians to compute matrix operations, such as those involved in regression, from linked microdata without sharing individual records (see, for example, Karr et al. 2009). Among the major limitations of this approach are that it relies on SRL, allows only datacustodians to analyse the microdata (i.e., non-data custodians cannot perform analysis) and that it is currently limited to a certain set of models. Alternatively, Kohnen and Reiter (2009) consider the novel problem of how data custodians, without sharing sensitive variables, can together produce synthetic linked microdata for public use. Limitations of this approach are that synthetic data can be time consuming to produce and that it can be hard to guarantee that the synthetic data do not distort some important relationships.

In contrast to the above approaches discussed in the literature, this article considers a more practical and straightforward approach to managing disclosure risk from linked microdata. In particular, this article considers the presence of a so-called Integrating Authority (IA) that is trusted to perform the following roles:

1. Link microdata collected by two data custodians.
2. Maximise the inherent utility or value of the linked microdata. This may include application of consistent standards and classifications, statistical editing and imputation.
3. Allow analysts to access the linked microdata in order to fit models.
4. Ensure the level of disclosure risk of the regression output is acceptable to the data custodians.

The IA is allowed to observe the microdata collected by the data custodians. The data custodians do not mask the microdata they provide to the IA in any way. The data custodians not only have access to the microdata they provided to the IA but, as analysts, they also have access to the regression output released by the Integrating Authority.

There are at least three benefits to an IA. First, the IA manages the complexity involved in linking microdata and managing disclosure risk – this is important since many data custodians do not have the specialised capability in, for example, standardising linking fields, editing and imputation, record linkage and data access. Second, since the IA observes the linking fields, it is possible to conduct a clerical review on the set of links and to refine the method of record linkage. This essential task appears impractical when linking fields are encrypted. Third, a more optimal trade-off between disclosure risk and the utility of the analysis is possible. With an IA, only the disclosure risk of the regression output needs to be managed; under the alternatives mentioned above, the disclosure risk must be managed from record linkage to construction of the regression output itself.

There are some major potential disadvantages of the IA framework. First, some data custodians may be prohibited by law, from disclosing information to any another organisation. This would mean the IA framework would not apply. Second, fulfilling the role of an IA is potentially a costly exercise. This may lead the IA to pass this cost burden onto analysts by charging a substantial fee for access. Moreover, it is the IA that decides how to fulfil its roles in any given situation. For example, the IA decides which variables to include on the linked microdata and how analysts will access the linked microdata (e.g., public use file or via a remote server, as discussed below). These decisions may suit some analysts but not others.

Once the record linkage step is completed by the IA, its next step is to facilitate access to the microdata. In this article, the IA releases regression output via a remote analysis server (see Reiter 2002, Gomatam et al. 2005, Sparks et al. 2008, Lucero and Zayatz 2010). A simple model for a remote server is:

1. An analyst submits a query, via the Internet, to the analysis server.
2. The analysis server processes the analyst's query on the linked microdata. The statistical output (e.g., regression coefficients) is modified or restricted in order to ensure the risk of disclosure is acceptably low.
3. The analysis server sends the modified output, via the Internet, to the analyst.

One key protection against disclosure afforded by remote analysis is that the analyst is restricted from viewing the microdata. However, an analyst may attempt to use the regression output to infer the value of variables on the linked microdata. Such attempts are commonly called *data attacks*. Once the value of these variables is inferred, the attacking

analyst can attempt one of the well-understood methods of disclosure (e.g., attribute disclosure through linkage); for a review see Shlomo (2007). The IA can provide analysts with disincentives to conducting attacks in the first place. For example, analysts could be required to sign confidentiality agreements to access to the remote server. If the agreement is violated by an analyst, access to the server could be revoked.

This article is about how an IA can manage the risk of a data custodian successfully attacking the linked microdata. A data custodian's attack would involve using the microdata it supplied to the IA and the regression output released by the remote server to infer the value of variables about a person or organisation that were collected by the other data custodian. Data custodians will commonly collect name and address, which means if such an attack is successful, the value of any variables that are inferred could be directly attributed to the person or organisation who provided that information. In other words, disclosure occurs automatically after a successful attack.

The problem of managing the disclosure risk of regression output released via a remote server has been the subject of significant recent attention in the literature. The literature on this problem focuses on the situation where there is a single data custodian responsible for managing access to its microdata (i.e., unlinked microdata). In the case of remote analysis for model fitting, most effort has been directed at linear regression. Gomatam et al. (2005) considered imposing restrictions to stop analysts reconstructing coefficients for a sensitive linear model, an example of which is a model with highly accurate predictions of a sensitive characteristic (see Bleninger et al. 2010 for an empirical investigation). Taking a completely opposite approach, Dwork and Smith (2009) describe the concept of *differential privacy*, which imposes no restrictions but instead relies on perturbation of statistical output alone to manage the disclosure risk. Many authors have considered imposing both restrictions and perturbation (e.g., Sparks et al. 2008); this article takes such an approach. One limitation of a remote server is that analysts are restricted to using the set of statistical analysis procedures that are supported by the remote server. This article only briefly mentions the more moderate disclosure risk of attacks made by noncustodians since, as mentioned, there is a considerable literature on this problem.

Section 2 describes how a data custodian may attack the linked microdata when the remote server naively releases standard regression output for models that are fitted to binary data. Section 3 proposes simple protections that an IA can implement in a remote server to reduce the success rate of these attacks. Section 4 evaluates the utility and disclosure risk of the proposed approach in a real situation and in a simulation. Section 5 makes some final comments.

2. Attacks Without Any Protection

This section describes how a data custodian can attack the linked microdata if the remote server naively releases standard regression output. Consider an IA linking microdata collected by two data custodians, referred to as A and T. Data Custodian A is the *attacker* and Data Custodian T is the *target*.

This article makes the assumption that all links between records are correct (i.e., each pair of records that are linked correspond to the same person or organisation) and that the name and address of all linked records are known to Data Custodian A. In practice, linkage

is rarely perfect and it is well known that errors arising during the linkage process provide some level of protection against disclosure (see Ch. 18 in Herzog et al. 2007). From the perspective of managing disclosure risk, the assumption that linkage errors do not arise is conservative.

Many authors distinguish between variables that are sensitive (e.g., income) and those that are not sensitive, where only the risk of disclosing sensitive variables needs to be managed. However, the legislation that guides how the Australian Bureau of Statistics, and many other data custodians, manages disclosure risk does not distinguish between sensitive and nonsensitive variables.

Let $\mathcal{D}$ be a set of records from the linked microdata comprising n records: a binary outcome variable y and a vector of K binary covariates $\mathbf{x}$. For the ith record, define $(y_i, \mathbf{x}_i)$ where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki}, \ldots, x_{Ki})'$ and $i = 1, \ldots, n$. Data Custodian T collected y and the K_T column vector $\mathbf{x}_T$ and Data Custodian A collected the K_A column vector $\mathbf{x}_A$ so that $\mathbf{x} = (\mathbf{x} = (\mathbf{x}_T', \mathbf{x}_A')')'$ and $K = K_T + K_A$. In other words, if we define $\mathbf{X} = (\mathbf{X}_T, \mathbf{X}_A) = (\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots \mathbf{x}_n)'$, Data Custodian T supplied $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)'$ and the $n \times K_T$ matrix $\mathbf{X}_T$ and Data Custodian A collected the $n \times K_A$ matrix $\mathbf{X}_A$. Therefore we may now write $\mathcal{D} = (\mathbf{y}, \mathbf{X})$.

An attack by Data Custodian A involves using regression output released by the remote server and $\mathbf{X}_A$ to infer the value of one or more elements of $(\mathbf{y}, \mathbf{X}_T)$. Therefore, for the purposes of this article, if a variable on the linked microdata is collected by both data custodians (e.g., a linking variable), it is defined as a covariate in $\mathbf{x}_A$.

In general, a good strategy for Data Custodian A's attack on a record involves ensuring $\mathbf{x}_A$, used in the calculation of the statistical output, uniquely identifies the target record on the linked microdata. This ensures there is 1–1 mapping between the target record's value of $\mathbf{x}_A$ and name and address. As Data Custodian A collected $\mathbf{X}_A$, this could readily be achieved.

Noncustodians present much less of a disclosure risk. Firstly, since they do not have access to $\mathbf{X}_A$, they can only use the statistical output released by the remote server in an attack. Secondly, even if an attack was able to reconstruct $(y_j, \mathbf{x}_j)$, attributing the jth record to a person or organisation is more difficult without name and address (see Skinner and Shlomo 2008).

Subsections 2.1, 2.2 and 2.3 describe attacks using standard regression output, including estimates of regression coefficients, estimates of their variance and test statistics, respectively.

2.1. Regression Coefficients

The standard estimate of the regression coefficient β for models fitted to binary variables (e.g., logistic regression, linear regression), denoted by $\hat{\beta}$, is obtained by solving the score equation

$$Sc(\beta; \mathcal{D}) = 0, \tag{1}$$

where $Sc(\beta) = \Sigma_i \mathbf{x}_i'(y_i - \mu_i)$ and $\mu_i = g(\mathbf{x}_i'\beta)$ for some link function g. It is well known that fitting a model to $\mathcal{D}$ is equivalent to fitting a model to the C counts contained in the vector $\mathbf{n}$, where $\mathbf{n} = \{n_c : c = 1, \ldots, C\}$ and n_c is the number of records belonging to the

cth pattern in $(y, \mathbf{x})$ (see McCullagh and Nelder 1989). As an aside, if y was instead a multinomial response with M categories, the appropriate score function would involve $M - 1$ equations of the form of (1). A multinomial response model fits into the framework developed here, but for simplicity we do not consider it further.

This section shows how Data Custodian A can attack $(\mathbf{y}, \mathbf{X}_T)$ – this involves using $\hat{\beta}$ and $\mathbf{X}_A$ in an attempt to infer the value of one or more elements of $(\mathbf{y}, \mathbf{X}_T)$.

2.1.1. Solving the Estimating Equations from a Single Model

Consider Data Custodian A substituting $\hat{\beta}$ into (1) and then attempting to solve for some elements of $(\mathbf{y}, \mathbf{X}_T)$. If the number of patterns in $\mathbf{x}_A$ is C_A, Data Custodian A's attack can exploit the following:

1. The K constraints imposed by $\hat{\beta}$ through (1)
2. Knowledge of $\mathbf{X}_A$
3. $(\mathbf{y}, \mathbf{X}_T)$ has only binary elements.

This attack can be as simple as conducting a grid search. Other more sophisticated search techniques could also be used. Of course this search could be more targeted if, for instance, Data Custodian T were to release frequency counts of y or $\mathbf{x}_T$ to the public. For example, the ABS, as potential Data Custodian T, releases frequency counts from its Census microdata after the counts have been perturbed by a small amount.

2.1.2. Solving Estimating Equations from Multiple Models

This attack involves fitting different models to the same set of data values in $\mathcal{D}$ (i.e., the same set of records and variables) by:

1. Changing the dependent variable
2. Changing the link function (e.g., linear, logistic and probit)
3. Transforming variables (e.g., creating an interaction term).

The regression coefficients for each fitted model impose additional constraints on $(\mathbf{y}, \mathbf{X}_T)$ via (1). The idea behind this attack is to impose sufficient constraints so that Data Custodian A can solve for one or more elements of $(\mathbf{y}, \mathbf{X}_T)$.

Example 1: Solving for all unknowns. Denote the data values in $\mathcal{D}$ by $\mathbf{Z} = (\mathbf{X}, \mathbf{y}) = (\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_n)'$, where z_{im} to be the mth element of $\mathbf{z}_i$. Consider Data Custodian A fitting the mth model where the outcome variable for the ith record is $y_i^{(m)} = z_{im}$ and the covariate for the ith record is $\mathbf{x}_i^{(m)}$, which is obtained by dropping z_{im} from $\mathbf{z}_i$. Denote the standard estimate of the regression coefficients from the mth model by $\hat{\beta}^{(m)}$ and denote $\hat{\mu}_i^{(m)} = g(\mathbf{x}_i^{(m)'}\hat{\beta}^{(m)})$. Data Custodian A's attack involves solving for one or more elements of $(\mathbf{y}, \mathbf{X}_T)$ given $\mathbf{X}_A$, $\hat{\beta}^{(m)}$ and the constraint

$$\Sigma_i \mathbf{x}_i^{(m)}(y_i^{(m)} - \hat{\mu}_i^{(m)}) = 0, \tag{2}$$

for $m = 1, \ldots, M$. Clearly, as M increases so does the number of constraints.

Example 2: Solving for unknowns in one estimating equation. Continuing *Example 1*, consider the lth estimating equation in (2) when Data Custodian A fits M models such that $y_i^{(m)} = y_i$ and the lth element of $\mathbf{x}_i^{(m)}$ is by definition x_{il}, for all $m = 1, \ldots, M$.

Further consider that Data Custodian A collected x_l so that it knows which $H = \Sigma_i x_{il}$ records contribute to the lth estimating equation. The constraint imposed by the lth estimating equation in (2) reduces to

$$\Sigma_{i,x_{il}=1} y_i - \Sigma_{i,x_{il}=1} \hat{\mu}_i^{(m)} = 0,$$

for $m = 1, \ldots, M$. This imposes M constraints on the $H \times (K_T + 1)$ unknowns for the H records contributing to the lth estimating equation. This number of unknowns could be considerably less than *Example 1*. An extreme example is when $H = 1$, which means there are $(K_T + 1)$ unknowns and M constraints.

Example 3: Imposing more constraints by creating a new variable. If Data Custodian A collected the variable t, it could repeat the attack in *Example 1* or *2* but where y_i is replaced with $y_{new,i} = y_i t_i$ for all i. By imposing the additional constraint $y_{new,i} = 0$ if $t_i = 0$, Data Custodian A can focus on solving y_i for *only* records with $t_i = 1$. This additional constraint could considerably reduce the number of unknowns.

2.1.3. Counts

Consider if Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$ and aims to infer $\mathbf{T} = \Sigma_i \mathbf{x}_i' y_i$, which are counts of y in the margins of $\mathbf{x}$. Given $\hat{\beta}$ and (1), this is straightforward since $\mathbf{T} = \Sigma_i \mathbf{x}_i' \mu_i$. The disclosure risks of frequency counts are well known (see, for example, Shlomo 2007). Counts of one would lead to disclosure. Counts of one can also be obtained through differencing, as discussed below.

2.1.4. Differencing

A standard differencing (see, for example, Gomatam et al. 2005; Shlomo 2007) attack involves fitting the same model to two sets of records that are identical except that one record is dropped from one of the sets. Data Custodian A can be sure only the target record is dropped if the dropping condition uniquely identifies the record and if it collected all the variables in the dropping condition. Differences in the estimated regression coefficients from the two models can be used in an attempt to infer the values of the dropped record's variables.

Example 4: Differencing attack by dropping a record. Consider if Data Custodian A wants to infer y_r, the value of y for rth record. Data Custodian A can fit a linear regression model with $\mathbf{x} = \mathbf{x}_A$ before and after dropping the rth record. Denote the value of the estimated regression coefficients before and after dropping the rth record by β_o and $\beta_{o(r)}$, respectively. Also denote $\mathbf{y}_{(r)}$ and $\mathbf{X}_{(r)}$ by $\mathbf{y}$ and $\mathbf{X}$ after removing the rth row, respectively. Since Data Custodian A knows β_o, $\beta_{o(r)}$, $\mathbf{X}_{(r)}$ and $\mathbf{X}$, it can calculate $\mathbf{S}_o = \mathbf{X}'\mathbf{X}\beta_o = \mathbf{X}'\mathbf{y}$ and $\mathbf{S}_{o(r)} = \mathbf{X}_{(r)}'\mathbf{X}_{(r)}\beta_{o(r)} = \mathbf{X}_{(r)}'\mathbf{y}_{(r)}$ and take the difference $\mathbf{S}_{o(r)} - \mathbf{S}_o = \mathbf{x}_r' y_r$. Since y_r is the only unknown, Data Custodian A can infer it directly.

2.1.5. Fishing

Fishing attacks involve fitting two models that are only different in one small way. Of interest is whether the two sets of coefficients are the same or whether they are different; how the coefficients change is not of interest. An example is given below.

Example 5: Fishing by slightly changing the definition of a variable. Consider linked microdata where Data Custodian A collected a variable for small area geography and age in single years and Data Custodian T collected a sensitive variable. Data Custodian A would know if there was one record in a particular small area with age equal to 100 years and may seek to infer the value of the sensitive characteristic for the record. Data Custodian A could fit two models to the records in the small area which are exactly the same, except that the first includes a binary covariate that takes the value one when age is between 40 and 100 *and* the sensitive characteristic is present and the second model includes a binary covariate that takes the value one when age is between 40 and 99 *and* the sensitive characteristic is present. If the regression coefficients from these two models are different, Data Custodian A infers that the 100-year-old has the condition; otherwise Data Custodian A infers that the 100-year-old does not have the condition.

2.2. *Estimated Variance of Regression Coefficients*

The estimated variance of $\hat{\beta}$ is $\widehat{Var}(\hat{\beta}; \mathcal{D}) = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}$, where $\hat{\mathbf{V}}$ is diagonal with ith element $v^{-1}\left(\partial\mu/\partial\eta\right)^2$ evaluated at $\mathbf{x} = \mathbf{x}_i$ and $\beta = \hat{\beta}$, v is the variance function for the model, and $\eta = \mathbf{x}'\beta$. Given $\hat{\beta}$, $\widehat{Var}(\hat{\beta}; \mathcal{D})$ can impose up to $K(K-1)/2$ constraints on $\mathbf{X}$. These constraints could be exploited to assist with an attack on estimated regression coefficients. Consider the simple linear regression model where $\widehat{Var}(\hat{\beta}; \mathcal{D}) = \hat{\phi}(\mathbf{X}'\mathbf{X})^{-1}$ which, after taking the inverse and multiplying by released dispersion parameter $\hat{\phi}$, gives the table of counts $\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_T'\mathbf{X}_T & \mathbf{X}_T'\mathbf{X}_A \\ \mathbf{X}_A'\mathbf{X}_T & \mathbf{X}_A'\mathbf{X}_A \end{pmatrix}$. Many of the attacks in Subsection 2.1 (e.g., differencing attacks and fishing) can be used against $\widehat{Var}(\hat{\beta}; \mathcal{D})$. They are not discussed further here.

2.3. *Other Statistical Output*

Regression analysis would normally include exploratory data analysis, use of test statistics and graphical plots to assess the model fit. Univariate and multivariate exploratory analysis involving binary variables will often involve frequency counts, which are well known to be a disclosure risk (see references below). Such work goes beyond the scope of this article, but will form the subject of future work.

Statistics used to assess model fit or goodness-of-fit (see Hosmer and Lemeshow 2000), say $t = t(\hat{\beta}, \mathcal{D})$, are functions of the microdata $\mathcal{D}$ and an estimate of β. Again, many of the attacks in Subsection 2.1 (e.g., differencing attacks and fishing) can be used against t. They are not discussed further here.

Graphical diagnostics are frequently used to assess model fit. The disclosure risk of plotting record-level values is high and has been considered by many authors (see O'Keefe and Good 2009 and O'Keefe et al. 2012). Consider if a remote server releases $\hat{\beta}$ and a plot which shows that the predicted value for a record is p. Given $\mathbf{x}$ has only binary elements, there will in general be only a single value of $\mathbf{x}$ such that $p = \mu(\mathbf{x})$. Furthermore, if the record has a unique value for $\mathbf{x}_A$ on the linked microdata, then Data Custodian A can infer $\mathbf{x}_T$ for the person about which the record relates.

3. Attacks in the Presence of Protections

This section proposes some simple protections against the attacks described in the previous section. The objective of these protections is to significantly reduce the likelihood of a successful attack while making a small impact on the utility of the analysis. Subsections 3.1 and 3.2 consider protections by imposing a general set of restrictions and by introducing uncertainty into regression coefficients, respectively. Subsection 3.3 considers attacks on estimated regression parameters in the presence of these protections. Subsections 3.4 and 3.5 describe protections for the variance of the estimated regression parameters and for diagnostic test statistics, respectively.

3.1. Protection: Imposing General Restrictions

Several restrictions are suggested below. These restrictions do not necessarily defend against a particular attack, but are designed to significantly hinder attacks while resulting in only a minor reduction in utility. When designing a set of restrictions to manage disclosure risk, it quickly becomes clear that a series of legitimate regression models *could* be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter. This challenge is discussed in detail by Cox et al. (2011).

Some analysts may have good reasons for fitting a model which is not permitted by the set of restrictions below. The IA could decide to relax some restrictions if: the analyst promises to fit a small number of predefined models (this could be verified by the IA); if the IA believes errors, such as incorrect or missed links, in the linked microdata provide substantial protection; or if the analysis has high utility. For obvious reasons, the IA would be more willing to relax restrictions for analysts who are not data custodians, as long as they promise not to share the regression output publically.

If the IA is not willing to relax one or more restrictions so that an analyst may fit a particular model, the IA may provide the analyst with an alternative mode of access to the linked microdata. One example would be for the IA to provide the analyst with the C counts required to fit the model, though the counts will almost certainly need to be carefully perturbed to manage the risk of disclosure.

Some possible general restrictions include:

(a) Limit the number of model covariates, K, by imposing the restriction that $K < 30$. Models with a large number of covariates may impose considerable constraints on unknowns. In very few cases would legitimate analysis be impacted by this restriction.
(b) Impose a minimum number of observations or covariate patterns by imposing the restrictions $n \geq 50$ and $C > 50$. This restriction aims to ensure a minimum number of unknowns. Remembering that C is the number of counts to which the model is fitted, $C \leq 2^K$ effectively means that $K > 5$.
(c) Adjusted $R - squared < 0.95$ (see also Gomatam et al. 2005). Other cut-off values can be considered. Inferential disclosure occurs when a model's prediction of a sensitive variable, y, is highly accurate and all covariates for the target record are known (e.g., $\mathbf{x} = \mathbf{x}_A$). This restriction is designed to prevent inferential disclosure. This protection will rarely be required since accurate predictions of binary outcomes

are rare. (Aside: inferential disclosure is fundamentally based on model assumptions. Some would argue that inferences which rely on model assumptions cannot lead to disclosure, because there is uncertainty about whether the model assumptions are true.)

(d) Each variable in the model must be non-zero for at least ten records. As all variables are binary this means $\Sigma_i x_{ik} \geq 10$ and $\Sigma_i(1 - x_{ik}) \geq 10$ for all k, $\Sigma_i y_i \geq 10$ and $\Sigma_i(1 - y_i) \geq 10$. This provides some protection against attacking a single estimating Equation (see *Example 2*) by ensuring there will be a minimum of ten unknowns.

(e) $(C - C_A) \geq 10K$. This ensures that there are ten times the number of unknown counts than there are constraints imposed by the estimating equation.

(f) New variables may only be created by multiplying two variables originally on the microdata as long as both variables were collected by the same data custodian. This restriction aims to prevent a data custodian from, almost arbitrarily, reducing the number of unknowns as in *Example 3*.

(g) Exclude variables from the linked file if they have limited analytic value. This limits the potential prior knowledge a data custodian can use in attacks. This decision must be made by the IA after consultation with potential analysts.

(h) Restrict variables which are naturally only useful as model covariates (e.g., marital status, age, sex, geography) from being dependent variables. This will hamper attempts to solve the estimating equation by changing the choice of dependent variable (see point 1 in Subsection 2.1.2). See also Gomatam et al. 2005 for another justification for this restriction.

It makes sense to impose data custodian-specific restrictions (e.g., see (e) above) because the disclosure risk naturally depends upon which data custodian is performing the attack. For data custodian-specific restrictions to make sense it must be assumed that there is restricted (e.g., to publications) sharing of regression coefficients between data custodians and that data custodians are aware of what regression coefficients they are able to share. What if this assumption is not realistic? The implication is that if one data custodian is restricted from fitting a model then all data custodians and non-data custodians must be restricted from fitting the model. In other words – *restriction for one means restriction for all*.

While the details are not within the scope of this article (for details see O'Keefe and Chipperfield 2013), the IA will need to decide what restrictions, if any, to place on subsetting records (i.e., defining the records in $\mathcal{D}$). If there is no restriction on subsetting, a data custodian may be able to arbitrarily target records to drop in differencing attacks. On the other hand, the flexibility of subsetting is very important since it allows analysts to make inferences about a specific population of interest.

If the number of models that are fitted is allowed to be arbitrarily high, the corresponding set of constraints may be such that an attacking data custodian can *solve the estimating equation*. Therefore it is worth mentioning a basic indicator of the risk of this attack succeeding. Consider when Data Custodian A fits its mth model to $C_{(m)}$ counts, where $C_{(m)}$ is the same as C but for the mth model and $C_{A(m)}$ is the same as C_A but for the mth model. Consider $L_A = \Sigma_m L_{A(m)}$, where $L_{A(m)} = C_{(m)}^{-1} C_{A(m)}$. The numerator of $L_{A(m)}$ is the number of constraints Data Custodian A can impose on the $C_{(m)}$ counts (see point 2 in

Subsection 2.1.1) to which the mth regression model was fitted. When $L_A > 1$ there are potentially more constraints than unknown counts, at which point the IA could perhaps audit the models fitted by Data Custodian A. Refining this indicator and developing similar indicators for other attacks would be an interesting line of future work.

3.2. Protection: Introducing Uncertainty into the Released Regression Coefficients

Two simple ways of introducing uncertainty into regression coefficients are now mentioned. The first protection is that a different random sample of records is dropped (see Sparks et al. 2008) for every distinct model that is fitted. Specifically, for each $k = 1, \ldots, K$, one record with $x_k = 1$ is randomly selected and dropped from $\mathcal{D}$. This means K records will be dropped. Denote $\mathcal{D}_{drop}$ to be $\mathcal{D}$ after dropping records in this way. Estimates of regression coefficients will not be biased by dropping records in this way, since it does not affect the distribution of y conditional on $\mathbf{x}$. As many applications involving linked microdata have a large number of records, dropping records in this way will generally only have a small impact on the accuracy of estimates. Note that dropping a completely random sample of records for every model fitted (see Sparks et al. 2008) provides limited protection in the present setting. Consider dropping 50 randomly selected records as a protection against the attack in *Example 2*, where $n = 50{,}000$ and $H = 5$ so that $x_k = 1$ for only five records. Since it is unlikely that $x_k = 1$ for any of the dropped records, it is equally unlikely that the attack in *Example 2* will be affected by dropping records in this way.

The second protection involves adding noise (for a review see O'Keefe and Chipperfield 2013) to the RHS of (1). Consider the estimator $\hat{\beta}^*$ of β, obtained by solving

$$Sc(\beta; \mathcal{D}_{drop}) = \mathbf{E}^*, \tag{3}$$

where the microdata used in the regression are $\mathcal{D}_{drop}$ not $\mathcal{D}$, $\mathbf{E}^* = (E_1^*, \ldots, E_k^*, \ldots, E_K^*)'$, $\mathbf{E}_k^* = \phi u_k^*$, ϕ is a scaling factor for the perturbation that needs to be set by the integrating authority and u_k^* s are independently generated variables from the uniform distribution on the range $(-1,1)$. Other distributions can be considered. The regression coefficients $\hat{\beta}$ are perturbed via $\mathbf{E}^*$. The value for ϕ is best determined through empirical investigation and simulation, which is discussed below. The distribution for u_k^* is bounded so that the impact of perturbation is bounded. The contribution of a record to the kth estimating equation is in the range $(-1,1)$, which is also the range of the perturbation, u_k^*. As many attacks attempt to uncover the values of variables for a single record, this is arguably a minimum degree of perturbation.

The distribution of the perturbations in $\mathbf{E}^*$ are independent so that $Var(\mathbf{E}^*)$ is a diagonal matrix. The joint distribution of $\mathbf{E}^*$ across different models should also be independent with an important exception: the same values of $\mathbf{E}^*$ should be used if exactly the same model is fitted. This condition stops estimation of $\hat{\beta}$ by fitting exactly the same model a number of times and averaging over the $\hat{\beta}^*$s.

3.3. Attacks Using the Released Estimated Regression Coefficients

Here we revisit the attacks of Section 2 in the presence of the protections mentioned above. It is assumed here that ϕ and the rules for dropping records are in the public domain.

3.3.1. Solving the Estimating Equation

Define $\hat{\beta}^{(m)*}$ to be the same as $\hat{\beta}^{(m)}$ except that it is obtained by solving (3) rather than (1). Consider solving the estimating equation in *Example 1* but where the regression parameter, $\hat{\beta}^{(m)*}$ instead of $\hat{\beta}^{(m)}$, is released. Define $\mathcal{D}^{(m)}_{drop}$ to be $\mathcal{D}$ after randomly dropping records for the mth model. Data Custodian A's attack now involves finding, over all possible subsets $\mathcal{D}^{(m)}_{drop}$ of $\mathcal{D}$, a unique solution for one or more elements of $\mathbf{y}$ given

$$-\phi 1 \leq \left\{ \Sigma_{i \in D^{(m)}_{drop}} \mathbf{x}^{(m)}_i (y^{(m)}_i - \hat{\mu}^{*(m)}_i) \right\} \leq \phi 1, \tag{4}$$

for $m = 1, \ldots, M$, where $\hat{\mu}^{*(m)}_i = g(\mathbf{x}^{(m)'}_i \hat{\beta}^{*(m)})$ and 1 is a K vector of 1s.

The protection provided by perturbation and dropping records depends upon the many possibly interacting factors implicit in (4). This makes it difficult to make any general conclusions about the protections they provide against disclosure. Clearly, the protection provided by perturbation is driven by ϕ. When looking at (4), it is clear that as ϕ increases the interval becomes wider and the probability of a unique solution (i.e., disclosure) becomes smaller. The method of dropping records would ideally prevent strict constraints being imposed on the terms in (4). If $y = 1$ for 99% of records, then an attack could assume, with high probability of being correct, that $y = 1$ for all dropped records. Making this assumption would impose a further constraint on the unknown values of y – in particular, if the first element of $\mathbf{x}$ was a constant, then the first element of $\mathbf{A} = \Sigma_{i \in D^{(m)}_{drop}} \mathbf{x}^{(m)}_i y^{(m)}_i$ in (4) would be constant over m. The first element of $\mathbf{A}$ could no longer be assumed to be constant if there was some uncertainty about how many records were dropped (e.g., instead of dropping one randomly selected record with $x_k = 1$, drop $1, 2, \ldots,$ or T randomly selected records with $x_k = 1$ with probability $1/T$).

3.3.2. Counts

Consider how $\hat{\beta}^*$ protects against estimating $\mathbf{T} = \Sigma_{i \in \mathcal{D}} \mathbf{x}'_i y_i$. If Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$, it can compute $\hat{\mathbf{T}}^* = \Sigma_{i \in \mathcal{D}} \mathbf{x}_i \hat{\mu}^*_i$, where $\hat{\mu}^*_i = g(\mathbf{x}'_i \hat{\beta}^*)$. Data Custodian A knows the minimum and maximum value for the counts in $\mathbf{T}$ are given by the corresponding elements of $\mathbf{T}_{min} = \hat{\mathbf{T}}^* - (\phi + K)\mathbf{1}$ and $\mathbf{T}_{max} = \hat{\mathbf{T}}^* + \phi \mathbf{1}$, respectively. The '$K$' in the expression for $\mathbf{T}_{min}$ reflects the fact that Data Custodian A knows that up to K records could be dropped from each estimating equation.

3.3.3. Differencing

Consider how perturbation protects against differencing attacks on counts (see *Example 4*), assuming for the moment that no records are randomly dropped (i.e., $\mathcal{D}_{drop} = \mathcal{D}$). Consider if Data Custodian A regresses $\mathbf{y}$ on $\mathbf{x} = \mathbf{x}_A$ before and after dropping the rth record. Accordingly define $\mathcal{D}_{(r)}$, $\mathbf{T}_{(r)} = \Sigma_{i \in \mathcal{D}_{(r)}} \mathbf{x}'_i y_i$, $\hat{\mathbf{T}}^*_{(r)} = \Sigma_{i \in \mathcal{D}_{(r)}} \mathbf{x}_i \hat{\mu}^*_{i(r)}$, where $\hat{\mu}^*_{i(r)} = g(\mathbf{x}'_i \hat{\beta}^*_{(r)})$, and $\hat{\beta}^*_{(r)}$ to be exactly the same as $\mathcal{D}$, T, $\hat{\mathbf{T}}^*$ and $\hat{\beta}^*$, respectively, except that they are computed after the rth record is dropped. Data Custodian A can compute an estimate of $\mathbf{x}'_r y_r$ by

$$\Delta_{(r)} = \hat{\mathbf{T}}^* - \hat{\mathbf{T}}^*_{(r)}. \tag{5}$$

If any element of $\Delta_{(r)}$ has magnitude greater than 2ϕ, Data Custodian A can infer that $y_r = 1$. It is also not hard to see that if $y_r = 0$ this differencing attack will never succeed. This means the success rate of this attack depends upon the probability that $y = 1$ for the target records. It is also not hard to see that, as K increases and ϕ decrease, the probability of this attack succeeding increases.

Now consider the same differencing attack but where records are randomly dropped, as discussed previously. Denote $\mathcal{D}_{(r)drop}$ to be the result of randomly dropping records from $\mathcal{D}_{(r)}$. Now $\hat{\beta}^*$ and $\hat{\beta}^*_{(r)}$ are calculated from $\mathcal{D}_{drop}$ and $\mathcal{D}_{(r)drop}$ instead of $\mathcal{D}$ and $\mathcal{D}_{(r)}$, respectively. Of course, the IA does not reveal which records are dropped so that $\mathcal{D}_{(r)drop}$ and $\mathcal{D}_{drop}$ are not known to Data Custodian A. Accounting for this uncertainty, it is easy to show if any element of $\Delta_{(r)}$ has magnitude greater than $2\phi + K$ (the difference between $\mathbf{T}_{min}$ and $\mathbf{T}_{max}$), Data Custodian A can infer that $y_r = 1$.

3.3.4. Fishing

Randomly dropping records as described above provides an effective protection against fishing attacks since, for every distinct model that is fitted, a different random sample of records is dropped. This will mean, continuing with *Example 5*, that the regression coefficients for the two models will be different whether or not the 100-year-old has the condition. Only if the same model is fitted repeatedly (i.e., the chosen link function, the set of records, and dependent and independent variables are all the same) should the same set of records be dropped. Otherwise this protection can be removed by averaging.

3.4. Variance of Estimated Regression Coefficients

Given the perturbation and model distributions are independent, the sandwich estimator for the variance of $\hat{\beta}^*$ is

$$\widehat{Var}\left(\hat{\beta}^*; \mathcal{D}_{drop}\right) = \widehat{Var}\left(\hat{\beta}; \mathcal{D}_{drop}\right) + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop} Var_*(\mathbf{E}^*)(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop}, \tag{6}$$

The first term in (6) is the estimated variance of the standard estimator of β obtained from solving (1), but based on $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. An analytic expression for the first term is $(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}_{drop}$, where $(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}$ is $\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}$ but based on $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. Alternatively the first term can be calculated from $\mathcal{D}_{drop}$ using the Bootstrap or Jackknife (see Chambers and Skinner 2003 p. 105). The second term in (6) measures the variation due to perturbation where it is easy to show, using the variance of the Uniform distribution, that $Var_*(\mathbf{E}^*)$ is diagonal with kth element $var_*(\phi u^*_k) = \phi^2/3$. The analyst can make valid inferences about β using $\hat{\beta}^*$ and (6), without knowing anything about the perturbation itself. It is interesting to note that the first and second terms of (6) are $O(n^{-1})$ and $O(n^{-2})$ respectively, which means that the impact of perturbation on variance is small.

Using the same reasoning as in Subsection 2.2, releasing (6), where the first term is computed analytically, would represent a high risk of disclosure. Instead consider computing the first term using the Jackknife. Denote θ as the analytic variance estimate of $\hat{\beta}$ and denote $\hat{\theta}$ as the corresponding Jackknife variance estimate of θ. The Jackknife estimate has a level of uncertainty due to the process, denoted by v, of allocating selection units to replicate groups. In particular, the coefficient of variation of $\hat{\theta}$ due to this process is $CV_v(\hat{\theta}) \approx 2(R-1)^{-1}$, where R is the number of replicate groups, $CV_v(\hat{\theta}) = Var(\hat{\theta})\hat{\theta}^{-2}$

(see Shao and Tu 1995, p. 196) and m/n is negligible. As long as R is not too large, this uncertainty in $\hat{\theta}$ will mask θ. This means that computing the first term in (6) using the Jackknife will mask the entire RHS of (6). For example, if the Jackknife standard error estimate is 0.2 and is based on $R = 50$, a 95% confidence interval for the estimate is $(0.31, 0.46)$.

It is difficult to see how (6) could be used in a differencing attack or be used to impose any constraint that would be useful to help solve the estimating equation. Since (6) is based on $\mathcal{D}_{drop}$ it is protected from fishing attacks. A further protection is to release only the diagonal elements of (6) so that only the variances of the regression coefficients are released.

3.5. Other Statistical Output

Given $\hat{\beta}^*$ instead of $\hat{\beta}$ is released, it makes sense that an analyst would be interested in $t^* = t(\hat{\beta}^*, \mathcal{D}_{drop})$ rather than t. The statistic t^* for the adjusted R^2, leverage, dispersion parameter and the Hosmer Lemeshow and chi-squared statistics will have their usual interpretation (i.e., replacing $\hat{\beta}$ and $\mathcal{D}$ with $\hat{\beta}^*$ and $\mathcal{D}_{drop}$ does not affect their interpretation).

Since $\hat{\beta}^*$ is not a likelihood estimator, it is not strictly valid for $\hat{\beta}^*$ to be used to evaluate likelihood-based diagnostic statistics. However, it is easy to show that it is approximately valid to do so in large samples. Standard likelihood-based test statistics (e.g., Likelihood Ratio Test and Deviance Test) involve evaluating the model log-likelihood $l(\hat{\beta}|\mathcal{D})$, where $\hat{\beta}$ is the standard ML estimator and $\mathcal{D}$ are the microdata used to fit the model. Using a second order Taylor Series approximation to $l(\beta|\mathcal{D})$ centred around $\hat{\beta}$ and noting $\hat{\beta}^* = \hat{\beta} + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{E}^*$, it follows that $l(\hat{\beta}^*|\mathcal{D}) \approx l(\hat{\beta}|\mathcal{D}) - 3^{-1}trace\{(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\}$ which means for large n that $l(\hat{\beta}^*|\mathcal{D}) \approx l(\hat{\beta}|\mathcal{D})$. Furthermore, if the number of dropped records is small then $l(\hat{\beta}^*|\mathcal{D}_{Drop}) \approx l(\hat{\beta}|\mathcal{D})$. For large n, this means that a standard likelihood-based test statistic evaluated at $\hat{\beta}^*$ and $\mathcal{D}_{drop}$ is approximately the same as a standard likelihood test statistic (i.e., $t^* \approx t$). This approximation is verified in empirical evaluations.

In small samples, it may be worthwhile to adjust some statistics to make them valid. For example, the standard Wald Test statistic is $t_W = \hat{\beta}'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\beta$ and is distributed as chi-squared with K degrees of freedom. The adjusted Wald statistic is

$$t_W^*(\hat{\beta}^*, \mathcal{D}_{drop}) = \hat{\beta}^{*\prime}\left[(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1} + (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1} Var(\mathbf{E}^*)(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})_{drop}^{-1}\right]\hat{\beta}^*,$$

and is chi-squared with K degrees of freedom.

The only protection of t^* from attacks is that it is calculated from $\mathcal{D}_{drop}$ rather than $\mathcal{D}$. To be consistent with the protections given to regression parameters (see Subsection 3.2), consider the perturbed statistic

$$t^{**} = t^* + e(t)u^*, \tag{7}$$

where $e(t)$ bounds the maximum influence that a record on the microdata has on the statistic t, and u^* is a random variable sampled from the uniform distribution on the range $(-1,1)$. If the same model is fitted then the same value for u^* must be generated (cf. averaging over $\mathbf{E}^*$ s). All the attacks discussed previously on regression parameters can be

reformulated to be attacks on diagnostic statistics, t^{**}. For reasons of space these are not mentioned.

For example, in the case of the dispersion parameter for a logistic regression, ideally $\hat{\phi}^* = (n-K)^{-1}\Sigma_i(y_i - \mu_i^*)^2 v^{-1}(\hat{\mu}_i^*)$ would be released. Since $e(\phi) \approx n^{-1}$, the released dispersion parameter $\phi^{**} = \phi^* + e(\phi)u^* = \phi^* + O(n^{-1})$, which means that perturbation will only have a small impact. Moreover, it is easy to show, using first-order Taylor Series approximations, that for many test statistics $t^{**} = t + O(n^{-1})$ – implying that the difference between the standard and released statistics will be small. This is verified in a limited empirical study.

For statistics used in hypothesis testing, only the ranged p-value for the test statistic, t^{**}, should be reported, rather than the value of the test statistic and the degrees of freedom. The degrees of freedom for t and t^{**} for the above mentioned test statistics are the same, using as justification the fact that $\phi^{**} \approx \phi$. Sparks et al. (2008) suggest reporting the p-values in the ranges $[0, 0.001)$, $[0.001, 0.01)$, $[0.01, 0.05)$, $[0.05, 0.1)$ and $[0.1, 1)$.

The challenge of confidentialising graphical output, including exploratory data analysis, in remote analysis systems is discussed by Sparks et al. (2008) and by many other authors (for a review see O'Keefe and Chipperfield 2013). This however, has not considered the risks from linked data. This is an interesting and useful avenue for future work.

4. Evaluation of Risk and Utility of a Remote Server: Linking the Australian Census to the Migrants Database

The ABS Census of Population and Housing provides economic and social information about migrants living in Australia. However, there are certain questions of great interest about migrants that the Census data alone cannot answer. One key question is how migrant visa class, assigned prior to arrival in Australia, is related to post-arrival social and economic outcomes. The different visa classes include *family, humanitarian, skilled, onshore* and *other*. Answering such a question is made possible through linking the Census with the Department of Immigration and Citizenships (DIAC) Settlement Database (SDB) which collects *visa class*. These answers would assist with the future development and evaluation of immigration programs and support services for migrants.

The Census 2006 microdata are made up of more than 20 million records. The reference period for the Census is 8 August 2006. For this study, the SDB had a reference period from 1 January 2000 to 8 August 2006 (Census night) and contained the records of 861,000 persons who, during that period, were granted visas to live permanently in Australia. DIAC provided the SDB to the ABS for the purpose of linking it with the Census. The variables used to probabilistically link records on the SDB and Census were *age* (in years), *month and day of birth, marital status* (five categories), *sex, country of birth, year of arrival, religion, main language* and *small area geography*. About 530,000 records were linked. For the purposes of this study, the linked file includes select Census variables, the SDB variable *visa class* and the linking variables *age, marital status, sex, country of birth, year of arrival, main language* and *small area geography*. For the purpose of this study we assumed that the linking variables *religion* and *month and day of birth* were not included on the linked data. This means DIAC would have access to seven variables and small area geographic information on the linked microdata. If more SDB

variables were included on the linked microdata, the disclosure risk would likely be greater than that measured below.

In this study the ABS is the IA and Data Custodian T and DIAC is Data Custodian A. The ABS, as an IA, is planning to release the SDB Census-linked microdata through its remote server. The ABS is legally obliged to ensure that the risk of disclosing information about a particular person is *unlikely*. This legislation (Census and Statistics Act 1905) does not distinguish between sensitive and nonsensitive variables and does not make a special provision for *trusted analysts*. (The case study here is an example of a general strategy of the ABS to link its Population Census to microdata collected by select government departments. Details on the legal framework behind an IA in Australia can be found on the ABS website).

Subsection 4.1 considers the utility of modelling with and without the protections of Section 3, and Subsection 4.2 considers the disclosure risk in a high-risk scenario.

4.1. *Empirical Evaluation of Utility*

While there are many possible research questions, one of particular importance to policy makers is to what extent migrants have difficulty finding employment after they arrive in Australia and how this is related to visa class. A useful way to answer such a question is to fit a regression model to employment with a range of covariates, including visa class. Tables 1 and 2 give the results of fitting such a model to two populations- the first is all migrants living in the Australian Capital Territory (ACT) and the second is all migrants living in the ACT who arrived after 2001, respectively. The set of restrictions of Subsection 3.2 did not prevent the models being fitted.

The results show that $\hat{\beta}^*$ with $\phi = 1$ (remembering that ϕ controls the magnitude of the perturbation) and the standard estimator $\hat{\beta}$ were very similar. As mentioned above, the standard errors of $\hat{\beta}^*$ can be computed by using either an analytic or Jackknife expression for the first term in (6). Tables 1 and 2 show that the difference between the two variance estimates is generally small and tends to be larger for coefficients of covariates that have a low frequency. Consequently, the tests for the statistical significance of the regression coefficients were almost identical whether they were based on $\hat{\beta}^*$ with Jackknife standard errors or $\hat{\beta}$ with analytic standard errors. The one exception was in Table 1, where the coefficient $55 < age < 64$ was not statistically significant at the 95% level after the protections were applied. Coefficients of covariates with a low frequency tend to be more influenced by perturbation of the score function. Tables 3 and 4 illustrate that the standard and released diagnostics statistics are very similar. Overall, this section illustrates that the protections had only a small impact on inference.

4.2. *Simulated Evaluation of Risks*

This section simulates attacks that could be conducted by an analyst with access to the SDB. The aim of such simulated attacks is to infer the value of one or more Census variables, using statistical output released by the remote server and the SDB. While the simulation does not involve use of the DIAC Census-linked microdata, it aims to replicate the possible attacks on the linked microdata. The benefit of simulation is that it is

Table 1. Impact of Protections on Regression Coefficients (ACT)

Variable name	Frequency ($n=$)	$\hat{\beta}$	$\hat{\beta}^*$	Analytic Standard Error of $\hat{\beta}$	Analytic Standard Error of $\hat{\beta}^*$	Jackknife Standard Error of $\hat{\beta}^*$
constant	5,161	−1.13	−1.06	0.19	0.20	0.19
school qual.	303	0.39	0.41	0.15	0.15	0.13
female	2,825	1.14	1.14	0.08	0.08	0.08
tertiary qual.	4,045	−0.56	−0.55	0.09	0.09	0.09
part-time student	493	−0.20	−0.20	0.14	0.14	0.14
full-time student	935	2.00	1.99	0.10	0.10	0.10
non-urban	25	−0.11	−0.11	0.60	0.61	0.79
not married	1,798	−0.39	−0.42	0.10	0.10	0.08
family visa	2,197	0.40	0.39	0.08	0.08	0.08
humanitarian visa	191	0.56	0.57	0.19	0.19	0.17
other visa	49	0.35	0.48	0.39	0.38	0.41
onshore visa	2,162	−0.19	−0.19	0.08	0.08	0.08
English spoken at home	1,324	−1.18	−1.21	0.15	0.15	0.17
English proficient	3,458	−0.75	−0.77	0.13	0.13	0.16
$25 \leq age \leq 34$	2,331	−0.08	−0.13	0.12	0.12	0.12
$35 \leq age \leq 44$	1,456	−0.05	−0.11	0.14	0.14	0.14
$45 \leq age \leq 54$	506	−0.24	−0.30	0.18	0.18	0.17
$55 \leq age \leq 64$	116	0.54	0.46	0.26	0.26	0.27

Table 2. Impact of Protections on Regression Coefficients (ACT and Year of Arrival Prior to 2001)

Variable name	Frequency ($n=$)	$\hat{\beta}$	$\hat{\beta}^*$	Analytic Standard Error of $\hat{\beta}$	Analytic Standard Error of $\hat{\beta}^*$	Jackknife Standard Error of $\hat{\beta}^*$
constant	1,529	−1.51	−1.42	0.46	0.50	0.56
school qual.	88	0.83	0.86	0.27	0.28	0.27
female	825	1.30	1.31	0.18	0.18	0.19
tertiary qual.	1,226	−0.75	−0.74	0.19	0.19	0.16
part-time student	156	−0.29	−0.31	0.29	0.29	0.28
full-time student	134	1.97	1.93	0.26	0.26	0.19
non-urban	11	0.92	1.49	1.14	1.37	0.99
not married	481	−0.45	−0.45	0.20	0.20	0.22
family visa	727	0.58	0.57	0.18	0.18	0.18
humanitarian visa	44	1.06	1.10	0.40	0.40	0.44
other visa	38	0.01	−0.17	0.53	0.55	0.80
onshore visa	795	0.10	0.10	0.16	0.16	0.17
English spoken at home	422	−1.05	−1.07	0.32	0.32	0.32
English proficient	1,032	−0.95	−0.95	0.30	0.30	0.29
$25 \leq age \leq 34$	583	−0.30	−0.39	0.33	0.33	0.36
$35 \leq age \leq 44$	587	0.09	0.00	0.34	0.34	0.33
$45 \leq age \leq 54$	181	−0.05	−0.14	0.39	0.42	0.38
$55 \leq age \leq 64$	27	0.68	0.39	0.59	0.61	0.66

Table 3. Impact of Statistical Disclosure Control on Diagnostic Statistics (ACT)

Statistic	Standard (t)	95% interval for t^{**}
Dispersion, ϕ	0.93	(0.92, 0.93)
$R-square$	0.18	(0.18, 0.18)
Likelihood Ratio	1052 (<0.001)	(1035, 1057) (<0.001)$^{\psi}$

$^{\psi}$only the ranged p-value is released.

possible to readily construct a situation that both is realistic and presents a high risk of disclosure.

The ABS, as an IA, would not reveal to data custodians which records were linked (e.g., in the Census-SDB linkage only 530,000 of the 861,000 SDB records were linked). However, it is assumed in this simulation that the attacker could identify a specific subpopulation of records that are very likely to be linked correctly. For example, in the Census-SDB linkage it may be inferred that certain subpopulations of records (e.g., proficient in English and high level of education) have a very high chance of reliably reporting linking variables, and so are likely to be linked correctly to their corresponding Census records.

4.2.1. Simulated Subpopulation

Assume the attacker fits models to a subpopulation of the linked microdata of size $n = 30$ or 50 records. This subpopulation could be defined in terms of small area geography, available on the SDB. Given the previous assumption, the attacker knows the exact set of records in the subpopulation. To make this simulation realistic, the attacker chooses to use eight variables on the linked microdata: small area geography to define the subpopulation of size n, the six other remaining SDB variables (see above), denoted by $\mathbf{x}$, and one Census variable (e.g., employment), denoted by y. In the notation of Section 2, the attacker knows $\mathbf{X}$ and seeks to infer y_i for some or all i. The variables for records in the subpopulation were independently generated 200 times in the following way:

- Each record has a unique covariate pattern in $\mathbf{x}$. Since $\mathbf{x}$ has dimension six, there are $2^6 = 64$ possible covariate patterns, of which $n = 30$ or 50 are randomly selected for the subpopulation.
- $S_y = \Sigma_i y_i = 3,6$ where y is generated from the logistic model $1/(exp(-\eta_i))$, $\eta_i = 1.6 + x_{1i} - 1.5x_{2i} + 1.3x_{3i} - 0.8x_{4i} + 1.3x_{5i} + 0.9x_{6i} + e_i$ and the e_i s are independent standard normal random variables. These model parameters were chosen arbitrarily but to be within the range of those in Tables 1 and 2 and to generate the desired range in S_y.

Table 4. Impact of Statistical Disclosure Control on Diagnostic Statistics (ACT and Year of Arrival Prior to 2001)

Statistic	Standard (t)	95% interval for t^{**}
Dispersion, ϕ	0.93	(0.91, 0.94)
$R-square$	0.18	(0.18, 0.19)
Likelihood Ratio	257 (<0.001)	(240, 261) (<0.001)$^{\psi}$

$^{\psi}$only the ranged p-value is released.

Since each value for x is unique and SDB contains the name and address for every record, disclosure automatically occurs if the attacker who has access to the SDB is able to infer the value of y_i for any $\mathbf{x}_i$. This is because there is a 1-1 correspondence between $\mathbf{x}_i$ and name and address for all i.

The ABS releases frequency counts from its Census microdata via its remote server. While a small amount of noise is added to these counts before they are released, it is frequently assumed in this simulation that S_y is in the public domain. This is a strong assumption since, as mentioned above, such counts are perturbed by a small amount.

In reality, it is unlikely all of the above conservative assumptions made for this simulation will be true. As a result, the disclosure risks would in reality be significantly lower than those measured in this section.

4.2.2. Attacks Using Regression Coefficients

The effectiveness of two attacks were measured on the 200 independently simulated subpopulations. It was interesting to see how the success of an attack was influenced by whether the remote server released:

- $\hat{\beta}$. This effectively means there is no (N) protection.
- $\hat{\beta}^*(\mathcal{D})$ computed from (3) but using $\mathcal{D}$ instead of D_{drop}. The protection is from perturbation (P) of the score function.
- $\hat{\beta}^*(\mathcal{D}_{drop})$ computed from (3). The protection is from perturbation and dropping a single randomly selected record (O,P), where O denotes dropping.

The first attack was *Solving the Estimating Equation* (SEE) (see *Example 1* and Subsection 3.3). When the remote server uses the O and P protections, SEE involved finding all possible values for $\mathbf{y}$ that are solutions to (4) given S_y, $\mathbf{X}$ and $\hat{\beta}^{*(m)}$ for $m = 1, \ldots, M$. Disclosure occurred for record j if, across all possible solutions, the value for y_j was always unique. Table 5 gives the proportion of SEE attacks that were successful in a range of scenarios. For example, Table 5 shows that when $n = 50$ and there were no protections, all values in $\mathbf{y}$ were disclosed in every one of the 200 simulated subpopulations from only a single model; if instead the P protection was used with $\phi = 1$, the success rate fell to 2%. A summary of the findings from Table 5 are described below.

- Releasing $\hat{\beta}$ was a high disclosure risk. The risk was 100% when y was the dependent variable.
- As ϕ increased the success rate reduced. However, the P protection on its own did not reduce the success rate to zero.
- The success rate increased as M, the number of fitted models, increased.
- The O protection on its own did not reduce the success rate.
- If only the P protection was used, uncertainty in S_y (see 6^{ψ} in Table 5) did not seem to provide much protection.
- If both the P and O protections were used, the disclosure risk was zero.

The second attack was Differencing Counts (DC) (see Subsection 2.1.3 and Subsection 3.3). The target record for a differencing attack was chosen completely at

Table 5. The Success Rate of Solving the Estimating Equation (SEE)

Number of Models	Dependent variable(s)	Defence[ψ/ψ]	n	S_y	Percentage of Attacks which inferred $y = 0$ for at least one record	Percentage of Attacks which inferred $y = 1$ for at least one record	Percentage of Attacks which inferred y for all records
1	y	N	30	6	100	100	100
1	x_1	N	30	6	93	42	10
1	y	N	50	6	100	100	100
1	x_1	N	50	6	82	9	3
1	y	O	30	6	100	100	100
1	y	O	50	6	100	100	100
1	y	P ($\phi = 1$)	30	6	16	0	0
1	x_1	P ($\phi = 1$)	30	3	92	12	2
1	x_1	P ($\phi = 1$)	30	6	76	12	0
1	x_1	P ($\phi = 1$)	50	3	86	4	0
1	x_1	P ($\phi = 1$)	50	6	44	0	0
3	y, x_1, x_2	P($\phi = 1$)	30	6	93	40	5
5	y, x_1, x_2, x_3, x_4	P ($\phi = 1$)	30	6	90	73	9
7	all	P ($\phi = 1$)	30	6	97	49	12
7[ψ]	all	P ($\phi = 1$)	30	6[ψ]	82	33	7
1	y	P ($\phi = 2$)	30	6	2	0	0
1	x_1	P ($\phi = 2$)	30	6	0	0	0
1	x_1	P ($\phi = 2$)	30	3	0	0	0
1	x_1	P ($\phi = 2$)	50	3	0	0	0
1	x_1	P ($\phi = 2$)	50	6	0	0	0
3	y, x_1, x_2	P($\phi = 2$)	30	6	0	0	0
5	y, x_1, x_2, x_3, x_4	P ($\phi = 2$)	30	6	43	8	2
7	all	P ($\phi = 2$)	30	6	69	78	2
3	y, x_1, x_2	O,P ($\phi = 1$)	30	3	0	0	0
5	y, x_1, x_2, x_3, x_4	O,P ($\phi = 1$)	30	3	0	0	0
7	all	O,P ($\phi = 1$)	30	3	0	0	0

[ψ] while $S_y = 6$ the attacker only knew $S_y = 5$, 6 or 7.

[ψ/ψ] N – No protection, O – Dropping one record completely at random, P – Perturbing.

Table 6. Differencing Attack

Defence	S_y	Success Rate (%)
N	30	100
P	30	5
O,P	30	0

random. Table 6 shows that the proportion of differencing attacks that were successful when protections N, P and (P and O) were used was 100%, 5% and 0% respectively.

For the results in Table 5, $L_{DIAC} = 0.5$ (see Subsection 3.1 where A = DIAC) for fitting a single model and $L_{DIAC} = 3.5$ when seven models were fitted. By contrast, for the models in Tables 1 and 2, $L_{DIAC} = 0.001$ and 0.002 respectively; these values are considerably smaller since most variables in the model were not SDB variables and the sample size was larger. Interesting further work would identify the optimal value for L_{DIAC} to trigger an audit by the ABS. If $L_{DIAC} > 1$ was to trigger such an audit, the audit would readily identify that the fitted models have the distinctive feature of the SEE attack (see Subsection 2.1.2). Remedial action could then be taken by DIAC and ABS to prevent further attacks.

The ABS, as an IA, could consider dropping variables from the linked microdata that are common to Census and SDB. If a common variable has limited analytic value, the ABS, as the IA, should consider dropping it from the linked microdata. This is particularly the case if a common variable is useful in uniquely identifying a record. Dropping such variables will limit the prior knowledge, and hence the effectiveness, of an attack.

5. Discussion

Modern advances have allowed vast amounts of microdata to be collected by data custodians. With increasing sophistication of policy makers and the consequent demand for more detail, linking such microdata across data custodians is becoming increasingly important. While the benefits to society of allowing access to linked microdata are significant, data custodians need to ensure that allowing access is unlikely to result in the disclosure of information about a particular person or organisation. The Australian Bureau of Statistics (ABS) is playing a lead role in developing a framework for the integration of Australian Commonwealth data. The role of an Integrating Authority (IA) is to maximise the inherent value of Commonwealth data to society, to facilitate access to the linked data and to ensure disclosure risk is acceptable. The ABS is developing infrastructure in the areas of record linkage and remote analysis to support its goal to become the lead IA in Australia.

This article proposes a set of protections that an IA can apply to statistical output from linked microdata. The evaluations show that the protections prevent disclosure in a high-risk scenario and have only a small impact on inferences for analysis involving moderate sample sizes. The method in the article can be readily extended to three or more data custodians. Importantly, this article shows that some popular protections against disclosure (e.g., dropping records, rounding regression coefficients or imposing restrictions on model selection) are perhaps not as effective as previously thought.

There is a need to extend the approach here to include analysis of continuous variables. Extensions to multilevel models is also important, since linked administrative data are often longitudinal in nature or contain a natural hierarchy.

6. References

Bleninger, P., Drechsler, J., and Ronning, G. (2010). Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study. Privacy in statistical databases, J. Domingo-Ferrer and E. Magkos (eds). New York: Springer.

Chambers, R.L. and Skinner, C.J. (2003). Analysis of Survey Data. Hoboken, NJ: John Wiley and Sons.

Churches, T. and Christen, P. (2004). Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4, Available at: http://www.pubmedcentral.nih.gov/tocrender.fcgi?iid=10563 (accessed June 2012).

Cox, L.H., Karr, A.F., and Kinney, S.K. (2011). Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think but not How to Act. International Statistical Review, 79, 160–183. DOI: http://dx.doi.org/10.1111/j.1751-5823.2011.00140.x

Dwork, C. and Smith, A. (2009). Differential Privacy for Statistics: What We Know and What We Want to Learn. Journal of Privacy and Confidentiality, 1, 135–154.

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Systems. Statistical Science, 20, 163–177. DOI: http://dx.doi.org/10.1214/088342305000000043

Herzog, T.N., Scheuren, F.L., and Winkler, W.E. (2007). Data Quality and Record Linkage. Berlin: Springer.

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression. Hoboken, NJ: John Wiley and Sons Inc.

Karr, A.F., Lin, X., Sanil, A.P., and Reiter, J.P. (2009). Privacy Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products. Journal of Official Statistics, 25, 125–138.

Kohnen, C. and Reiter, J.P. (2009). Multiple Imputation for Combining Confidential Data Owned by Two Agencies. Journal of the Royal Statistical Society Series A, 172, 511–528. DOI: http://dx.doi.org/10.1111/j.1467-985x.2008.00574.x

Lucero, J. and Zayatz, L. (2010). The Microdata Analysis System at the U.S. Census Bureau. Privacy in Statistical Databases, J. Domingo-Ferrer and E. Magkos (eds). New York: Springer.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models (2nd ed.). London: Chapman and Hall.

O'Keefe, C. and Chipperfield, J.O. (2013). A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. International Statistical Review. DOI: http://dx.doi.org/10.1111/insr.12021

O'Keefe, C. and Good, N. (2009). Regression Output from a Remote Analysis System. Data & Knowledge Engineering, 68, 1175–1186. DOI: http://dx.doi.org/10.1016/j.datak.2009.06.009

O'Keefe, C., Sparks, R., McAullay, D., and Loong, B. (2012). Confidentialising the Output of a Survival Analysis in a Remote Analysis System (to appear). Journal of Privacy and Confidentiality, 4, 127–154.

Reiter, J. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. Journal of Official Statistics, 18, 511–530.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Hoboken, NJ: John Wiley and Sons.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review, 75, 199–217. DOI: http://dx.doi.org/10.1111/j.1751-5823.2007.00010.x

Skinner, C. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-Linear Models. Journal of American Statistical Association, 103, 989–1001. DOI: http://dx.doi.org/10.1198/016214507000001328

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., and Keighley, T. (2008). Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics™. Computer Methods and Programs in Biomedicine, 91, 208–222.

Received January 2013
Revised October 2013
Accepted November 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 147–161, http://dx.doi.org/10.2478/jos-2014-0008

The Relative Impacts of Design Effects and Multiple Imputation on Variance Estimates: A Case Study with the 2008 National Ambulatory Medical Care Survey

Taylor Lewis[1], *Elizabeth Goldberg*[1], *Nathaniel Schenker*[1], *Vladislav Beresovsky*[1], *Susan Schappert*[1], *Sandra Decker*[1], *Nancy Sonnenfeld*[1], *and Iris Shimizu*[1]

The National Ambulatory Medical Care Survey collects data on office-based physician care from a nationally representative, multistage sampling scheme where the ultimate unit of analysis is a patient-doctor encounter. Patient race, a commonly analyzed demographic, has been subject to a steadily increasing item nonresponse rate. In 1999, race was missing for 17 percent of cases; by 2008, that figure had risen to 33 percent. Over this entire period, single imputation has been the compensation method employed. Recent research at the National Center for Health Statistics evaluated multiply imputing race to better represent the missing-data uncertainty. Given item nonresponse rates of 30 percent or greater, we were surprised to find many estimates' ratios of multiple-imputation to single-imputation estimated standard errors close to 1. A likely explanation is that the design effects attributable to the complex sample design largely outweigh any increase in variance attributable to missing-data uncertainty.

Key words: Health survey; missing data; item nonresponse; fraction of missing information.

1. Background

The National Ambulatory Medical Care Survey (NAMCS) has been administered by the National Center for Health Statistics (NCHS) since 1973. While aspects of the sample design and survey instrument have evolved over the past twenty-five years, its objective has always been to collect and disseminate nationally representative data on office-based physician care. The ultimate sample unit is a doctor-patient encounter, drawn systematically from the terminus of a multistage, clustered sample design. Like many other surveys, the NAMCS is not immune to the potentially detrimental effects of missing data. As Figure 1 demonstrates, the (unweighted) item nonresponse rate for patient race, one of the most analyzed demographics, increased appreciably between 1999 and 2008. Such nonresponse on race has been experienced in the context of other NCHS health care surveys as well. For example, Kozak (1995) found that hospitals participating in the National Hospital Discharge Survey underreported race to varying degrees.

[1] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A
Email: tlewis@survey.umd.edu

Acknowledgments: The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention. The authors thank the Editor, Associate Editor, and referees for comments that helped to improve the article.

Groves et al. (2002, Sec. 1.2) cited three issues that can arise with missing data due to nonresponse: (1) biases in point estimators; (2) inflation of the variances of point estimators; and (3) biases in customary estimators of precision. In this article, we focus on the third issue, and in particular the extent to which multiple imputation (Rubin 1987) results in estimates of precision that differ from those under single imputation in the context of the NAMCS with missing data on race.

Variance estimates for situations such as ours have been explored by Li et al. (2004), who used a bootstrap re-imputation scheme adapted to complex surveys (Shao and Sitter 1996) to account for missing-race uncertainty in the 2000 NAMCS. Li and her colleagues observed a few instances where the bootstrap re-imputation suggested standard errors should be inflated by up to 30%, but concluded most estimates necessitated an inflation of 6% or less. Their findings quelled concerns for a while, but as one can infer from Figure 1, the item nonresponse rate for race in the 2000 NAMCS was roughly half where it stood in 2008.

This article reports on research conducted at NCHS, using data from the 2008 NAMCS, to assess whether multiple imputation would better reflect the missing-data uncertainty than single imputation, which is currently used in the NAMCS, in light of the recent nonresponse rates of about 30% on race. Using a model-based imputation method with predictors similar to those used in the 2008 NAMCS cell-based procedure, we compared results under multiple imputation to those under single imputation, and we found that the increase in the estimated standard errors with multiple imputation tended to be small. We concluded that the extremely large design effects (Kish 1965) for estimates involving race tended to transcend the additional missing-data uncertainty that would be reflected by multiple imputation. This is discussed with the help of some basic theory partitioning the overall estimated variance increase into a component attributable to the complex survey design and a component attributable to missing-data uncertainty.

Section 2 of the article provides an overview of the NAMCS sample design and describes the imputation method used in our study. In Section 3, we present the major results from the comparison of multiple imputation with single imputation. Section 4 concludes the article with a brief discussion pointing out limitations and suggestions for further research.

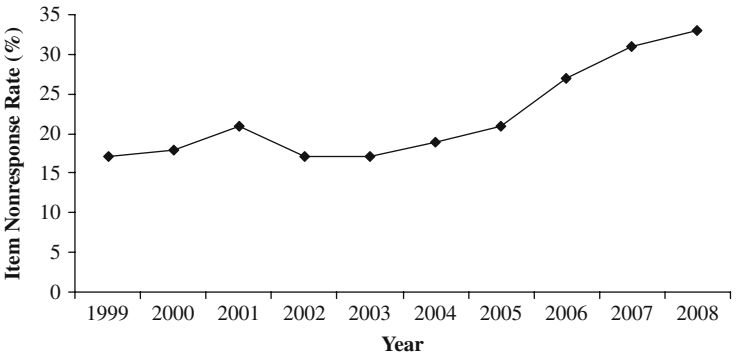

Fig. 1. Patient Race Item Nonresponse Rate Trend in the National Ambulatory Medical Care Survey, 1999 – 2008.

2. Data and Methods

2.1. NAMCS Sample Design

As previously noted, the NAMCS employs a multistage, clustered sample design. The primary sampling units (PSUs) consist of either single or grouped counties (or their equivalent), derived from a probability subsample of 112 PSUs from the 1985–1994 National Health Interview Survey (NHIS) design period. Within these PSUs, lists of non-federally employed physician practices obtained from the American Medical Association and American Osteopathic Association are stratified into fifteen specialty groups. A sample of physician practices is then selected from each stratum and randomly allocated into 52 subsamples, each corresponding to a week within the data collection period, the calendar year.

NCHS contracts with the U.S. Census Bureau to collect the patient visit information from sampled practices. Prior to their assigned one-week collection period, field representatives (FRs) meet with the physician or, more commonly, the physician's administrative staff, and analyze the expected count of pending patient visits. Based on this information, a systematic sampling interval is determined and utilized such that approximately thirty visits are selected over the course of the week. FRs try to recruit and train office staff to collect the sampled visits' data in real time, but more than half of the patient record forms (PRFs) are filled out by the FR using maintained patient files after the weeklong data collection period has concluded.

According to the 2008 NAMCS public-use data file documentation (NCHS 2009), a total of 3,319 physicians were selected, of whom 1,090 were ruled ineligible. Aside from having retired, common causes for ineligibility include a physician practicing in an institutional setting or as part of an emergency department outpatient facility. Of the 2,229 eligible physicians, 1,334 were contacted and agreed to participate, although 201 saw no patients during the data collection period randomly assigned. In the end, data were collected for 31,146 distinct visits. This number includes data from a supplemental sample of community health centers (CHCs) drawn with assistance from the Health Resources Services Administration and the Indian Health Service, of which a portion involved visits to non-physicians (e.g., nurse practitioners). Non-physician visits are excluded from the public-use file, which explains why the number of visits contained in the 2008 NAMCS public-use file (28,741) is fewer than analyzed in this article.

To compensate for the differential patient visit selection probabilities and physician-level nonresponse, a four-step weighting procedure yielded a final set of weights that can be used to better represent the target population. For more details on the weighting process, refer to Section I.K of NCHS (2009).

In addition to unit nonresponse caused by the fact that not all sampled physicians participate, the NAMCS is subject to item nonresponse in the returned PRFs. Some variables are more susceptible to missingness than others. Whereas most items' nonresponse rates are less than five percent, Section I.I.3 of NCHS (2009) lists specific rates for variables where the item nonresponse rate exceeds that threshold. Patient race has one of the highest rates: Of the 31,146 visits in the 2008 NAMCS, it is unknown for 10,149, or 32.6%.

The PRF extracts ethnicity and race from the physician records in accordance with the two-item format standardized by the U.S. Office of Management and Budget (1997).

The first item records whether the patient is Hispanic or Latino. Regardless of the response to the first, the second is a mark-all-that-apply with five races listed. A typical categorization for analysis breaks responses into six groups, cases where one and only one race was selected and a catch-all for individuals identifying with two or more races. Although we did investigate the imputation models' impact on the first question and the six racial categorizations, many are rare and yielded unstable estimates and standard errors. Because of this and for brevity purposes, we report a simplified, three-level race breakout: patients identified as white only, black only, or any other race (whether singly or in combination with white or black).

2.2. *Imputation Methods*

In this section we discuss the cell-based method used to impute missing race in the 2008 NAMCS, and contrast it with a model-based procedure that we felt was better suited to quantify the additional uncertainty reflected by multiple imputation. We also detail how we accounted for features of the complex sample design using this model-based approach.

The single-imputation method used in 2008 was based on a SAS® macro developed by Valverde and Marsteller (2007) that imputes missing race using a hybrid approach falling somewhere between a hot- and cold-deck (Andridge and Little 2010) and what Kalton and Kasprzyk (1986) term *hierarchical* imputation. When race is missing, the macro works dynamically to search for a donor on up to twenty-five matching criteria. For instance, the first criterion is to select a patient race randomly from a pool of donors within the same survey year, three-digit diagnosis code (see Section II.A.28 of NCHS 2009), and patient ZIP code. If no match can be found, the macro seeks a record of the same diagnosis code and patient ZIP code, but from the previous year's data.

Simply running the macro more than once to generate multiple imputations would not be prudent, since it ignores the imputation model's uncertainty. Rubin (1987) terms such a procedure *improper* (pp. 112–128). Rubin and Schenker (1986) offer the *approximate Bayesian Bootstrap* (ABB) as a way to perform proper multiple imputation in the cell-based setting. The ABB is akin to independently drawing a set of regression parameters from the posterior predictive distribution of an explicit imputation model prior to drawing each set of imputations. It was not immediately evident, however, what effect the hierarchical nature of the imputation macro would have on the theory underlying the ABB. We considered applying a bootstrap re-imputation scheme of the sort proposed by Efron (1994) and adapted to complex survey designs by McCarthy and Snowden (1985) and Shao and Sitter (1996), in the spirit of analyses undertaken by Li et al. (2004). In the end, we deemed a model-based multiple-imputation procedure most directly amenable to quantifying the increase in estimated variance in transitioning from single to multiple imputation.

The model-based procedure, sequential regression multivariate imputation (Raghunathan et al. 2001), was implemented using IVEware (http://www.isr.umich.edu/src/smp/ive/), free SAS-callable software developed by the Institute for Social Research at the University of Michigan, capable of imputing continuous, semicontinuous, categorical, and count variables. It uses an iterative algorithm which cycles through the variables with missing data, imputing the missing values of each variable conditional on the other

variables (Raghunathan et al. 2001). By imputing each variable in turn using those that came before or after, it builds interdependence among the data. Another useful feature is the ability to bound imputations within a specified range, something utilized in this and other NCHS imputation projects (e.g., Schenker et al. 2011).

In determining which covariates to include in the model-based procedure, we began by incorporating those utilized in the cell-based procedure and, based on input from subject matter experts, added variables we anticipated would help explain the missing data pattern and race itself, including patient age, sex, urban/rural indicator based on metropolitan statistical area (MSA), physician specialty group, reason for visit, natural logarithm of time spent with physician, and an indicator of who entered data into the PRF.

In addition to as many known covariates as possible, Rubin (1996) asserts imputations should be conditional on the sample design: "Minimally, major clustering and stratification indicators and sample design weights (or estimated propensity scores of being in the sample) should be included in imputation models" (p. 478). Indeed, a simulation by Reiter et al. (2006) exposes severe biases that can result from excluding such indicator variables when they explain the underlying mean function, even if the missingness mechanism is fully captured.

Nearly all the matching criteria in the cell-based method are at a finer level than PSU (i.e., ZIP codes generally lie within PSU boundaries). For the model-based method, we tried to include stratum and PSU indicators and sample weights as prescribed, but encountered convergence issues for the logistic regression parameters that did not cease until the PSU indicators were omitted. Reiter et al. (2006, p. 148) warn of such a problem:

> In some surveys the design may be so complicated that it is impractical to include dummy variables for every cluster. In these cases, imputers can simplify the model for the design variables, for example collapsing cluster categories, or including proxy variables (e.g., cluster size) that are related to the outcome of interest.

As a compromise, we incorporated local race distribution information from the U.S. Census Bureau's American FactFinder tool (http://factfinder2.census.gov/main.html). Specifically, we created two variables to house Census 2000 estimated proportions of non-Hispanic whites and non-Hispanic blacks at the ZIP code tabulation area level. For a portion of the cases (roughly 10%), patient ZIP code was unavailable. Where possible, we substituted physician practice ZIP code. For the remaining 3% of cases without a patient or physician ZIP, the race distribution variables were imputed, using IVEware's bounding feature to ensure proportions remained within [0, 1]. Kozak (1995) used a similar method at the county level, reporting: "Although not exact, the population distribution of a county appeared useful as a general indicator of the racial distribution of discharges from a hospital in the county" (p. 4).

2.3. *Multiple-Imputation Inferences*

In this section we introduce notation and formulas pertinent to inferences from multiply-imputed data as well as a few related metrics facilitating comparisons to singly-imputed data. Instead of a missing value being filled in once, multiple imputation calls for a missing value to be imputed M times ($M \geq 2$). In our study with the 2008 NAMCS, $M = 5$.

Each of the $m = 1, \ldots, M$ completed (observed plus imputed) datasets are analyzed individually and a particular quantity and its variance can be estimated through Rubin's (1987) straightforward combination rules given below.

If we let $\hat{Q}_m$denote the m^{th} completed-dataset estimate of a quantity Q, the quantity's overall multiple-imputation estimate is simply the average of the M estimates, $\bar{Q}_M = \frac{1}{M}\sum_M \hat{Q}_m$.

Let $\overline{U}_m$ denote the m^{th} completed-dataset estimated variance for $\hat{Q}_m$. The multiple-imputation estimated variance is the average of the M completed-dataset variances, $\overline{U}_M = \frac{1}{M}\sum_M \overline{U}_m$, plus a term reflecting the between-imputation variance of the estimate, $B_M = \sum_M \frac{\left(\hat{Q}_m - \bar{Q}_M\right)^2}{M-1}$.

After a finite imputation correction factor $\left(1 + \frac{1}{M}\right)$ is applied to the between-imputation variance component, the overall multiple-imputation variance formula is given by

$$T_M = \overline{U}_M + \left(1 + \frac{1}{M}\right) B_M. \tag{1}$$

A useful metric with a simple interpretation is the ratio of a quantity's multiple-imputation estimated standard error to its average single-imputation counterpart,

$$R = \sqrt{T_M / \overline{U}_M}. \tag{2}$$

The degree to which R exceeds 1 represents the percent increase in the estimated standard error attributable to *multiple* imputation.

Another related quantity is the *fraction of missing information* (FMI) (Rubin 1987, sec. 3.3; Wagner 2010), which can be approximated by the between-imputation variance component over the total variance,

$$FMI_{approx} = \left(1 + \frac{1}{M}\right) B_M / T_m. \tag{3}$$

Although the FMI typically depends to some extent on the percent of observations missing, it also depends on the analysis of interest and the extent to which the imputation model is predictive of the missing values. For example, if the imputation model is highly predictive, the FMI will tend to be substantially smaller than the item nonresponse rate.

3. Results

In an attempt to gauge the magnitude of missing-data uncertainty unaccounted for by single imputation, we calculated the ratio of multiple-imputation to average single-imputation estimated standard errors – Equation (2) in Subsection 2.3 – across a multitude of domains. For brevity, we present results from only a subset of those domains: the overall race distribution and the distribution by United States region, age group, and whether the patient has been diagnosed as diabetic. The estimated standard error ratios and other statistics related to these estimates are tabulated in Appendix.

The ratios for all domain estimates are plotted against their respective percent of observations missing in Figure 2. Most ratios exceed 1.0 only slightly, and just two surpass 1.1. These figures are in line with what was reported by Li et al. (2004), despite the patient race item nonresponse rate nearly doubling since the 2000 NAMCS data analyzed therein.

Intuition might lead one to expect the standard error ratios to increase with a higher item nonresponse rate. However, the plot exhibits no such trend. At least for the data at hand, the percent of missing observations alone does not predict the increase in estimated standard errors after multiply imputing. Estimates subject to 30% or more missingness are apparently no more severely underestimating the missing-data uncertainty by singly imputing than estimates with less than 30% missingness.

We followed numerous leads to explain the phenomenon, but most proved futile. For instance, we hypothesized the lopsided distribution of race might have triggered a software glitch. However, other than convergence issues discussed in Section 2, we concluded that IVEware performed soundly. As we will now discuss, the most reliable determinant of a small standard error ratio was found to be a large design effect in the underlying estimates.

Kish (1965, p. 193) defines a design effect as the ratio of the estimate's variance incorporating the complex design to the variance under a simple random sample of the same size

$$deff = \frac{\text{var}_{complex}(\hat{Q})}{\text{var}_{SRS}(\hat{Q})}. \tag{4}$$

The quantity we report in this article could perhaps more aptly be termed the *misspecification effect*, as it is the estimated variance accounting for the complex design features (i.e., stratification, clustering, and weights) over the estimated variance ignoring those features. Nonetheless, because these two terms are often colloquially exchanged for one another, we retain the more frequently utilized phrase.

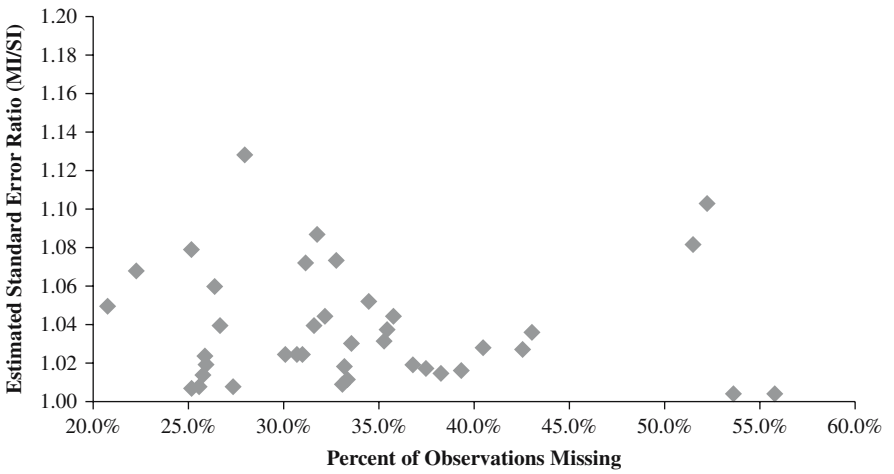

Fig. 2. The Relationship between the Percent of Observations Missing and the Ratio of Multiple-imputation (MI) to Average Single-imputation (SI) Estimated Standard Errors for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

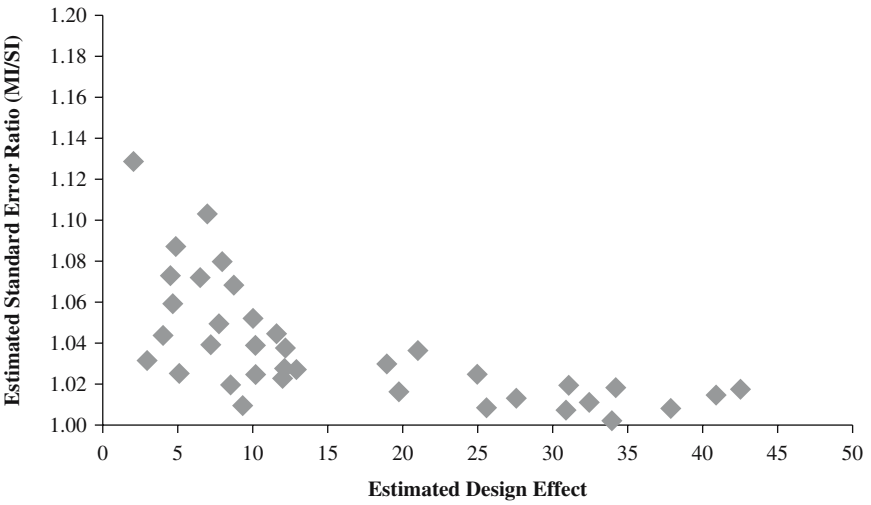

Fig. 3. The Relationship between the Average Completed-dataset Estimated Design Effect and the Ratio of Multiple-imputation (MI) to Single-imputation (SI) Estimated Standard Errors for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

Figure 3 illustrates the inverse relationship between the estimated standard error ratio and estimated design effect for the 2008 NAMCS data. Estimates with a larger design effect are clearly associated with a smaller increase in estimated standard error after multiple imputation. In one of Reiter et al.'s (2006) simulations a similar observation was made, where the multiple-imputation estimated standard error, even in the presence of a 30% item nonresponse rate, was only slightly larger than the complete data estimated standard error (i.e., the estimated standard error that would be obtained in the absence of nonresponse). The authors reason that the complex design "makes the within-imputation variance a dominant factor relative to the between-imputation variance. That is, the fraction of missing information due to missing data is relatively small when compared to the effect of clustering" (p. 146). Figure 3 demonstrates this concept over a range of design effects, using real data. Note that the x-axis scale was truncated at a design effect of 50 to allow for a clearer visualization of the patterns we wished to highlight. Although the truncation omits the two data points in the Appendix with the largest design effects – 70.38 and 97.34 – it does not substantively alter any of our observed patterns and conclusions. (A similar truncation is applied in Figure 4.)

Mentioned previously, an alternative gauge of missing-data uncertainty is the FMI (Wagner 2010). In fact, reproducing Figure 3 with FMI_{approx} of expression (3) on the vertical axis (not shown here) tells the same story. As the design effect increases, FMI_{approx} tapers. This occurs because the two metrics are monotonically related – our ratio of estimated standard errors is $\left(1 - FMI_{approx}\right)^{-\frac{1}{2}}$.

To further elucidate the relative impact of the design effect we can partition the increase in estimated variance into two components, that attributable to the complex sample design and that attributable to missing-data uncertainty as measured by using

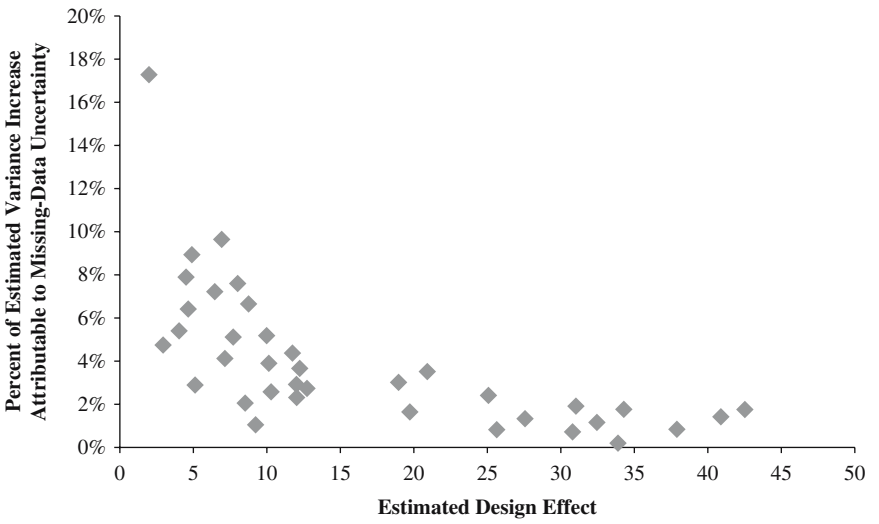

Fig. 4. The Percent of the Estimated Total Variance Increase Attributable to Missing-Data Uncertainty as a Function of the Average Completed-dataset Estimated Design Effect for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

multiple rather than single imputation. Specifically, we can conceptualize the term $\overline{U}_M$ in Equation (1) as being the product of $\overline{U}_{M(SRS)}$, the average completed-data-set variance assuming simple random sampling, and *deff*. Therefore, the approximate variance increase due to both the complex design and missing-data uncertainty can be written as

$$\Delta_M = \overline{U}_{M(SRS)} * (deff) + \left(1 + \frac{1}{M}\right) B_M - \overline{U}_{M(SRS)}. \tag{5}$$

The proportion of Δ_M attributable to missing-data uncertainty is simply the between-imputation term over the increase, or $\left(1 + \frac{1}{M}\right) B_M / \Delta_M$, whereas the proportion attributable to the complex design is the complement about 1, or $1 - \left(1 + \frac{1}{M}\right) B_M / \Delta_M$. We acknowledge, however, that this might not account perfectly for the two sources of increase, because the complex sample design could also affect the between-imputation term, B_M.

Figure 4 demonstrates the relationship between the design effect and the percent of the variance increase attributable to missing-data uncertainty as measured by multiple imputation. The pattern mirrors that appearing in Figure 3. In the presence of a larger design effect, the variance increase is dominated by the component attributable to the complex sample design. The figure suggests that, despite item nonresponse rates often exceeding 30%, a design effect of 10 or greater limits the impact of missing-data uncertainty to generally no more than 5% of the overall variance increase. Put another way, the portion of variance attributable to the complex design in these settings is at least 95% / 5% = 19 times greater than the portion attributable to missing-data uncertainty.

Another way to evaluate the relative increase in estimated variance due to the use of multiple rather than single imputation is to consider what the relative increase would be if the design effect were equal to 1. To approximate the answer, we fitted Lowess smoothers (Cleveland 1979), not shown here, to the data in Figure 3 with a variety of bandwidths that were large enough to avoid major jaggedness in the fitted curves. Extrapolating the curves to a design effect of one suggested a ratio of multiple-imputation to single-imputation estimated standard errors in the range of 1.08 to 1.1. Since, as mentioned earlier, the ratio equals $(1 - FMI_{approx})^{-\frac{1}{2}}$, it follows that the suggested range for FMI_{approx} is 14% to 17%. This range is consistent with a nonresponse rate of about 30% and an imputation model that is partially, not fully, predictive of the missing values.

4. Discussion

In this article, we presented results from a case study in which we evaluated the potential impact on estimated variances if a multiple-imputation strategy were adopted to handle instances of missing patient race in the 2008 NAMCS. The NAMCS sample design involves features such as clustering and highly variable analysis weights that result in extremely large design effects for estimates involving race. In these settings, we found multiple imputation increased estimated variances only modestly. Revisiting our key analytic quantity, the ratio of estimated standard errors in Equation (2), we can reason that as M goes to infinity, the ratio can be rewritten as

$$R \approx \sqrt{1 + \frac{B_M}{\overline{U}_M}}. \tag{6}$$

With a large design effect, the within-imputation component, $\overline{U}_M$, tends to be large relative to the between-imputation component, B_M, pulling the ratio towards 1.

At least among the domains investigated, the item nonresponse rate itself was not found to be predictive of the increase in estimated variance after multiply imputing the missing data. Even when the percent of imputed observations tops 30%, a large design effect can render multiple-imputation estimated standard errors only slightly greater than their single-imputation counterparts. For this reason, together with the increased complexity that multiple imputation poses to the typical NAMCS data user, it was decided to maintain a single-imputation approach for the NAMCS for the time being.

Despite the growing class of techniques available to compensate for missing data, the best way to handle nonresponse is to design data collection protocols preventing it from occurring in the first place (Lohr 1999). In mid-2009, NCHS raised FR awareness of the increased patient race item nonresponse rate, stressing the demographic's importance for analyses. The intervention appears to have been effective, as the item nonresponse rate for race dropped to 24% in the 2009 NAMCS and to 23% in the 2010 NAMCS. Albeit still high by many standards, at least the trend in Figure 1 has begun to reverse course.

Our study is not without limitations. For one, the domains analyzed herein are coarse in nature. It seems plausible that design effects may be attenuated for racial distributions

estimated for finer domains, which could produce scenarios where the proportionate increase in estimated variance due to using multiple imputation is larger than is reflected in this study.

Another limitation is that focus was restricted to only one variable, despite the fact that the NAMCS collects data on hundreds of other variables pertaining to the visit. In addition to feedback from NAMCS data users that patient race is a frequently utilized demographic, as previously mentioned, it is also subject to one of the highest item nonresponse rates. Although not presented here, we investigated another variable of key analytic interest, time spent with the physician, which was also susceptible to a high level of item nonresponse (26%) in the 2008 NAMCS. Similar findings were observed. Due to large design effects in the domains analyzed, multiple imputation increased estimated standard errors only slightly. As noted on page 18 of NCHS (2009), the item nonresponse rate for most other variables is 5% or less, so these are naturally of less concern.

A final limitation, noted in Subsection 2.2, is that we used a "compromise" method to reflect the features of the complex sample design in our imputation model. Had we accounted for those features perfectly, our results might have changed somewhat. However, we believe that our case study demonstrates an actual phenomenon for multiple reasons. First, variables related to the design features were included in the model. Second, as mentioned in Section 3, our case study yields results consistent with simulations reported in Reiter et al. (2006). Finally, if the survey clustering were more fully reflected in the imputation model, a likely result would be imputed values that are more differentiated, that is, less homogeneous, across the clusters. This might very well increase the design effects for each dataset completed by imputation, which, all else being equal, would accentuate the phenomenon displayed by our case study. Development of methods for reflecting design features parsimoniously in imputation models, such as by using random effects, is an important area for future methodological research.

Recent changes to the NAMCS sample design may prompt a re-evaluation at some point in the future. Beginning with the 2012 NAMCS, PSUs are no longer comprised of geographically clustered units. Instead, the universal list of physician offices is stratified by state and a sample selected within each, so the physician office now serves as the PSU. To the extent this new sample design alters the variability of weights or the heterogeneity of PSUs with respect to patient race, the magnitude of the design effects could change.

Aside from more empirical analyses such as the one discussed in this article, a simulation study and further theoretical research could foster a better understanding of the relationship between the design effect and the between-imputation component of variability reflected by multiply imputing missing data. Of particular interest would be to determine if and how the relationships we observed are moderated by how predictive the imputation model is and/or by alternative patterns of nonresponse.

Appendix

Appendix. Select Point Estimates, Estimated Standard Errors, Estimated Design Effects, and Indicators of Missing-Data Uncertainty from 2008 NAMCS Data Singly Imputed (SI) and Multiply Imputed (MI) by the Model-Based Method

Domain	Race	MI Estimate (%)	Average SI Estimated Standard Error (%)	MI Estimated Standard Error (%)	Estimated Standard Error Ratio (MI/SI)	Estimated Standard Error Assuming SRS (%)	Estimated Design Effect[1]	Percent Obs. Imputed (%)
Overall								
	White	84.6	1.243	1.265	1.018	0.212	34.283	33.2
	Black	10.5	0.949	0.961	1.013	0.181	27.618	25.8
	Other	4.9	0.837	0.851	1.017	0.128	42.538	37.5
Region								
Northeast	White	90.0	2.030	2.079	1.024	0.406	25.053	30.1
	Black	5.9	1.433	1.456	1.016	0.322	19.788	39.4
	Other	4.0	0.776	0.829	1.068	0.263	8.720	22.3
Midwest	White	88.9	1.316	1.349	1.025	0.412	10.195	30.7
	Black	9.0	1.285	1.314	1.023	0.371	12.014	25.9
	Other	2.1	0.341	0.352	1.031	0.198	2.978	35.3
South	White	82.5	2.203	2.217	1.007	0.396	30.903	25.2
	Black	15.5	2.192	2.197	1.002	0.376	33.965	18.8
	Other	2.0	0.211	0.238	1.128	0.148	2.037	28.0
West	White	81.3	3.943	3.956	1.003	0.470	70.380	55.8
	Black	5.3	0.710	0.783	1.103	0.269	6.937	52.3
	Other	13.4	4.007	4.024	1.004	0.406	97.340	53.7
Age								
Less than 15	White	79.5	3.200	3.261	1.019	0.574	31.087	36.8
	Black	13.5	2.965	2.989	1.008	0.481	37.939	27.4
	Other	7.0	1.318	1.354	1.027	0.369	12.771	42.6
15-24	White	81.5	2.400	2.420	1.009	0.785	9.351	33.1
	Black	13.7	1.946	2.042	1.049	0.699	7.743	20.8
	Other	4.8	1.512	1.555	1.028	0.434	12.126	40.5

Appendix. Continued

Domain	Race	MI Estimate (%)	Average SI Estimated Standard Error (%)	MI Estimated Standard Error (%)	Estimated Standard Error Ratio (MI/SI)	Estimated Standard Error Assuming SRS (%)	Estimated Design Effect[1]	Percent Obs. Imputed (%)
25-44	White	82.4	1.538	1.598	1.039	0.482	10.172	31.6
	Black	11.6	1.190	1.213	1.019	0.408	8.517	26.0
	Other	6.0	1.059	1.099	1.037	0.303	12.223	35.5
45-64	White	86.1	1.307	1.365	1.044	0.382	11.715	32.2
	Black	9.7	0.927	1.000	1.079	0.328	7.997	25.2
	Other	4.2	0.979	1.008	1.030	0.225	18.961	33.6
65-74	White	87.5	1.189	1.276	1.073	0.558	4.532	32.8
	Black	8.8	1.029	1.090	1.059	0.476	4.674	26.4
	Other	3.7	0.650	0.678	1.044	0.325	4.004	35.8
75 +	White	89.6	1.493	1.570	1.052	0.471	10.053	34.5
	Black	6.6	0.875	0.897	1.025	0.386	5.143	31.0
	Other	3.8	1.316	1.363	1.036	0.287	20.971	43.1
Diabetes								
No	White	85.0	1.262	1.276	1.011	0.221	32.507	33.4
	Black	10.0	0.950	0.957	1.008	0.188	25.613	25.6
	Other	5.0	0.858	0.870	1.014	0.134	41.006	38.3
Yes	White	81.4	1.842	1.975	1.072	0.723	6.501	31.2
	Black	13.9	1.685	1.751	1.039	0.628	7.188	26.7
	Other	4.7	0.957	1.041	1.087	0.433	4.895	31.8

Note:
[1]Average estimated design effect of the $M = 5$ completed datasets.

5. References

Andridge, R. and Little, R. (2010). A Review of Hot Deck Imputation for Survey Non-Response. International Statistical Review, 78, 40–64, DOI: http://www.dx.doi.org/ 10.1111/j.1751-5823.2010.00103.x.

Cleveland, W. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829–836, DOI: http://www.dx.doi.org/10.1080/01621459.1979.10481038.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, 89, 463–475, DOI: http://www.dx.doi.org/10.1080/01621459.1994.10476768.

Groves, R., Dillman, D., Eltinge, J., and Little, R., (Eds.) (2002). Survey Nonresponse. New York, NY: Wiley.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Kish, L. (1965). Survey Sampling. New York, NY: Wiley.

Kozak, J. (1995). Underreporting of Race in the National Hospital Discharge Survey. Advance Data from Vital and Health Statistics, No. 265. Hyattsville, MD: National Center for Health Statistics.

Li, Y., Lynch, C., Shimizu, I., and Kaufman, S. (2004). Imputation Variance Estimation by Bootstrap Method for the National Ambulatory Medical Care Survey, Proceedings of the Survey Research Methods Section of the American Statistical Association.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Brooks/Cole.

McCarthy, P. and Snowden, C. (1985). The Bootstrap and Finite Population Sampling. Vital Health Statistics, 2(95). Hyattsville, MD: National Center for Health Statistics.

National Center for Health Statistics (2009). 2008 NAMCS Micro-Data File Documentation. Division of Health Care Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Hyattsville, MD, Available online at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc08.pdf (accessed January 2014).

Office of Management and Budget (1997). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, Federal Register 62FR58781-58790. Available at: http://www.gpo.gov/fdsys/pkg/FR-1997-10-30/pdf/97-28653.pdf (accessed January 2014).

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, 27, 85–95.

Reiter, J., Raghunathan, T., and Kinney, S. (2006). The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data. Survey Methodology, 32, 143–150.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley.

Rubin, D. (1996). Multiple Imputation After 18 + Years (with discussion). Journal of the American Statistical Association, 91, 473–489.

Rubin, D. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366–374, DOI: http://www.dx.doi.org/10.1080/01621459.1986.10478280.

Schenker, N., Borrud, L., Burt, V., Curtin, L., Flegal, K., Hughes, J., Johnson, C., Looker, A., and Mirel, L. (2011). Multiple Imputation of Missing Dual-Energy X-Ray Absorptiometry Data in the National Health and Nutrition Examination Survey. Statistics in Medicine, 30, 260–276, DOI: http://www.dx.doi.org/10.1002/sim.4080.

Shao, J. and Sitter, R. (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278–1288, DOI: http://www.dx.doi.org/10.1080/01621459.1996.10476997.

Valverde, R. and Marsteller, J. (2007). A Revised Matching Routine for Imputing Missing Race and Ethnicity in the National Ambulatory Medical Care Survey, Unpublished internal manuscript of the National Center for Health Statistics.

Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. Public Opinion Quarterly, 74, 233–243, DOI: http://www.dx.doi.org/10.1093/poq/nfq007.

Received November 2011
Revised January 2014
Accepted January 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 163–166, http://dx.doi.org/10.2478/jos-2014-0009

Book Review

Books for review are to be sent to the Book Review Editor Jaki S. McCarthy, USDA/NASS, Research and Development Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.
Email: jaki_mccarthy@nass.usda.gov

Jennifer Madans, Kirsten Miller, Aaron Maitland, and Gordon Willis (Eds). *Question Evaluation Methods: Contributing to the Science of Data Quality*. Hoboken, NJ: John Wiley & Sons, Inc. 378 pp. 2011. Paperback: ISBN 9781118037003, price USD 64.20. E-pub: ISBN 9781118036983, price USD 52.99.

Website Q-Bank: http://wwwn.cdc.gov/qbank/

This book grew out of an interdisciplinary workshop on question evaluation methods and has as its goal to bring together knowledge from leading experts across different methods. The book consists of seven sections, or rather seven extended chapters, as each section contains a primary chapter describing a specific method and one or two shorter discussion chapters.

The first section opens with an excellent overview by Jack Fowler on behavior coding as a tool for evaluating survey questions. This chapter describes how behavior coding of interviewers and respondents is done and presents empirical evidence of its significance. It concludes with an outline of how this method should be fitted into question evaluation protocols and presents a well chosen reference list including key references in this field. The Fowler chapter provides an introduction to the novice in behavior coding and a good summary for those who have some experience with question evaluation. The two response chapters are aimed at the more advanced researcher. Nora Cate Schaeffer and Jennifer Dykema present a conceptual framework on how the interaction between respondent and interviewer affects data quality. They also present two very interesting summary tables (Tables 3.1 and 3.2) on the empirical associations between interviewers' and respondents' behavior and measurement quality. They then introduce the reader to conversation analysis as a tool and illustrate this with excerpts from the Wisconsin longitudinal study. Alisu Schoua-Glusberg broadens the discussion and focuses on the sociocultural context in which the survey interview takes place. Her remarks concern behavior coding as an evaluation tool, but are equally worthwhile for other question evaluation methods, such as cognitive interviews. As international, cross-cultural, and multilingual studies take on a greater importance in the modern world (cf. Harkness et al. 2010), researchers should realize that respondents will have different degrees of familiarity with the survey process;

in developing and pretesting survey questions, differences in communication styles in different cultural groups should be taken into account.

The second section on cognitive interviewing opens with a chapter by Kristen Miller, who gives a theoretical review of development of cognitive interviewing and describes a new integrative paradigm for question development and testing. This chapter is clearly intended for the knowledgeable and methodologically interested reader and is not meant as an introduction to cognitive interviewing. Those seeking such an introduction should read Willis's (2005) book first. The following two chapters by Gordon Willis and by Fred Conrad are critical rejoinders, and like the Miller chapter their discussion is aimed at cognitive survey methodologists.

Question evaluation and cognitive survey methodology are often seen as more qualitative approaches, but the next sections prove that this view is incorrect. Statistical modeling provides us with powerful tools for investigating measurement error and evaluating questionnaires. These quantitative methods are used in phase two of questionnaire evaluation. In phase one, the questionnaire is developed and pretested using more qualitative approaches, for example, expert evaluation and cognitive interviewing. The improved questionnaire is then implemented in an actual survey, ideally a field test.

In their chapter (Section 7) Brian Harris-Kojetin and James Dahlhamer describe what field tests are and the importance of collecting additional data, such as interviewer feedback. They illustrate this with examples from US federal statistical surveys. Field tests are fairly common in daily survey practice. Less common are the use of Multi-Trait Multi-Method (MTMM) matrices and specific experiments to collect data for quantitative questionnaire evaluation. Section five is devoted to split-sample experiments as a tool for collecting data for question evaluation. Jon Krosnick starts with a brief review of the experimental method and provides several insightful examples of methodological experiments on question wording, formats, and context. These are relatively large-scale field experiments aimed at quantitative analysis. Johnny Blair adds to this an outline for a more qualitative cognitive interview experiment. Theresa de Maio and Stephanie Wilson expand on this by emphasizing the importance of integrating a qualitative and quantitative approach. To quote: "this mixed-method approach allows us to understand what survey questions are actually measuring, and make better decisions about which questions to field".

Section six on the multitrait-multimethod approach deals with a special kind of experimental setup and its analysis. In his introductory chapter, Duane Alwin first describes the MTMM design as an approach to systematically collect data and gives a historical overview starting with the work of Campbell and Fiske in psychology and of Andrews in sociology and survey research. The concepts of reliability and validity in MTMM and in classical test theory are clearly explained and trait validity versus construct validity is discussed. Special attention is given to the role of memory in MTMM designs and recent applications of the MTMM approach. This chapter is a mixture between data collection and data analysis; data collected according to an MTMM design are by default analyzed using a Structural Equation Modeling (SEM) approach. In his response to Alwin, Peter Mohler gives an extended example of an MTMM study from the European Social Survey.

After the collection of large-scale quantitative data, be it through a regular survey, a specially designed field test, or a specific (experimental) design, there are various

statistical methods that provide powerful tools for quantitative questionnaire evaluation. For instance, the previously mentioned SEM approach can also be used to carry out multigroup comparisons and investigate measurement equivalence across different cultural or national groups (Vandenburg and Lance 2000; Hox et al. 2010). In cases where multiple items are used to measure one well-defined construct, Item Response Theory (IRT) is a promising analysis tool. Section three opens with a brief but informative chapter on Item Response Theory (IRT) and how it can be applied to questionnaire evaluation. IRT modeling focuses on scales that measure an underlying construct, using multiple items and a strict psychometric model. Bryce B. Reeve introduces the principle of IRT and illustrates how it can be applied in evaluating and refining questionnaires. He then introduces Computer Adaptive Testing (CAT) where a combination of qualitative pretest methods, such as expert evaluation and cognitive testing, and quantitative data analysis is used to produce an item bank with IRT-calibrated items. In CAT a respondent is then presented with an item in the middle range and an estimate is made of the person's scale score based on the response; then another item based on this estimate is selected from the item bank, and the process is repeated until the desired precision is reached. CAT allows for short questionnaires, adapted to the person's ability, with the desired precision. This is illustrated with PROMISS (Patient-Reported Outcomes Measurement Information System). In his rejoinder, Ronald Hays provides the reader with additional examples of the use of IRT in question evaluation. He also emphasizes that IRT analysis is extremely useful for detecting problematic items and building libraries of well-performing items, but that qualitative methods are needed to understand why an item performs badly. The next rejoinder by Clyde Tucker et al. is less a discussion of IRT and more an introduction to Latent Class Analysis (LCA) as a tool for questionnaire evaluation. In LCA an attempt is made to find an underlying latent categorical nominal or ordinal variable (latent classes) that explains the relationship between a number of observed variables. This is well illustrated with an example where LCA is used to classify respondents into good, fair, and poor reporters of expenditures.

Latent Class analysis (LCA) is then further introduced in section four by Paul Biemer and Berzofsky. In their conclusion they state that LCA is challenging for a novice. Their chapter proves them right; it requires more advanced statistical knowledge than the other chapters. Biemer and Berzofsky present the reader with a statistical introduction to LCA, its assumptions, and how to handle some common statistical problems. Together with the examples in the previous chapter it gives a good impression of how LCA can be used in questionnaire evaluation and in discovering response tendencies. In her rejoinder, Frauke Kreuter summarizes a comparison of different traditional questionnaire testing techniques (e.g., expert evaluation) and LCA; she also offers good guidelines on how test material and analysis results should be incorporated into question banks. Finally, Janet Harkness and Timothy Johnson go beyond LCA analysis as such, addressing issues in question design and pretesting that are somewhat neglected in general discussion, such as context effects.

Reasons to buy this book: Renowned experts from different disciplines introduce and discuss qualitative and quantitative methods of questionnaire evaluation. The methods introduced go beyond standard question evaluation methods such as expert evaluation and cognitive interviewing and focus on the collection and analysis of quantitative data for questionnaire evaluation. The quality of the contributions is high. The book is

accompanied by the very worthwhile Q-bank website (http://wwwn.cdc.gov/qbank/). Q-bank goes beyond more traditional question banks, providing the reader with an online database of questions that have been evaluated as well as their accompanying question evaluation reports.

Reasons not to buy this book: Although the book aims at a wide audience, not all chapters are easily accessible. Due to the format, a large introductory chapter followed by shorter rejoinders, the discussion aims at experts in the field. It is the well-edited proceedings of a multidisciplinary workshop and still reads as such.

In sum: I am glad I have read the book. I will certainly use (parts of) it in teaching advanced courses in survey methodology and it is a good accompaniment to the well-known earlier book by Presser et al. (2004). Both books should be in the library of survey researchers and statisticians in the private sector, government and academia, and the library of my institute now has both. However, if you have to advise a master or graduate student with limited monetary resources and have to choose one, I would recommend the book by Presser et al. as introduction.

References

Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.-E., and Smith, T.W. (2010). Survey Methods in Multinational, Multiregional, and Multicultural Contexts. New York: Wiley.

Hox, J.J., De Leeuw, E.D., and Brinkhuis, M.J.S. (2010). In Survey Methods in Multinational, Multiregional, and Multicultural Contexts, J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.P. Mohler, B.-E. Pennell, and T.W. Smith (eds). New York: Wiley.

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

Vandenberg, R.J. and Lance, C.E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices and Recommendations for Organizational Research. Organizational Research Methods, 3, 4–69.

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

Edith de Leeuw
Department of Methodology and Statistics
Faculty of Social and Behavioral Sciences
Utrecht University
PO Box 80.140
NL-3508 TC Utrecht
the Netherlands
Telephone: +31 (0)30 253 4438
E-mail: e.d.deleeuw@uu.nl

VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 167–169, http://dx.doi.org/10.2478/jos-2014-0010

Erratum

Erratum concerning the article "Are They Really Too Busy for Survey Participation? The Evolution of Busyness and Busyness Claims in Flanders" by Anina Vercruyssen, Bart Van de Putte, and Ineke Stoop published in Journal of Official Statistics, Volume 27, Number 4, 2011, pp. 619–632.

The unusually high odds ratio for the SCV 2002 dummy variable in Model 2 of Table 5 in the article is caused by the age of the 2002 survey respondents being missing from the merged data file with the data of the three surveys – an unfortunate error. The statistical analysis with the correct data file shows that the models actually provide even better support for our hypotheses. In contrast to Table 5 in the article, the effects of free time on week/workdays on busyness claims are robust (Table 5). We now also find significant and robust effects for claims of temporary busyness (Table 6), whereas the old table did not have any effects. In other words, there is stronger support from the data that respondents' doorstep statements on time pressure are true.

Corrected version of pages 627-629

Is this co-occurring decrease in leisure time and increase statements of (temporary) busyness coincidental, or is there truth behind the time concerns of respondents? Table 5 shows that those respondents who have less free time on work/week-days are indeed significantly more likely to have busyness claims and claims of temporary busyness, even when controlling for the interviewer effects, employment status (as an indicator of objective busyness), socio-demographic variables and possible interviewer effects. As for the interviewer effects, none of the variance components were significant (Table 5, Table 6). Both Table 5 and 6 also show that respondents who have a paid job are significantly more likely to make busyness statements and statements of temporary busyness. These results are in line with the literature on time and combination pressure: Those with a job are those who can experience combination pressure alongside to time pressure. These results show that the opportunity cost hypothesis and the bad timing hypothesis seem to apply for the SCV surveys.

5. DISCUSSION

The aim of this study was to determine whether the proclaimed increase in time and combination pressure in Western societies affects survey participation by investigating busyness claims ("too busy", "have no time") and statements of temporary busyness ("come back at another time") as statements made to decline survey participation. We found that these busyness related doorstep reactions increased significantly since 2002 in the investigated SCV surveys in Flanders (APS, 2002; 2005; 2007) and that the use of such reactions seems to be associated with a higher likelihood of also being a final refuser in these Flemish surveys. Moreover, we found that there is truth to these busyness claims: respondents with less free time are significantly more likely to state they are too busy or have no time, even after controlling for other indicators of time and combination pressure

Table 5. Two-level logistic regression for predicting busyness claims with objective indicators of busyness, controlling for interviewer effects

	Model 1				Model 2			
Busy	Odds ratio	Sig.	Random component	Sig.	Odds ratio	Sig.	Random component	Sig.
Level 1 predictors								
Intercept	0.076	***	2.162	n.s.	0.079	***	1.762	n.s.
Free time work-day	0.947	***	0.007	n.s.	0.979	*	0.026	n.s.
Free time non-work day	1.014	n.s.	0.015	n.s.	1.003	n.s.	0.022	n.s.
Paid job					1.358	***	0.451	n.s.
Age					1.005	n.s.	0.000	n.s.
Sex					1.273	n.s.	0.137	n.s.
Cohabiting					0.888	n.s.	0.835	n.s.
Children					1.042	n.s.	0.543	n.s.
Level 2 predictors								
SCV 2002					1.098	n.s.		
SCV 2005					1.077	n.s.		
N level 1	3552				3552			
N level 2	297				174			

Note: $*p \leq 0.05$, $***p \leq 0.001$; n.s. = not significant.

Table 6. Two-level logistic regression for predicting claims of temporary busyness with objective indicators of busyness, controlling for interviewer effects

	Model 1				Model 2			
Temporarily busy	Odds ratio	Sig.	Random component	Sig	Odds ratio	Sig.	Random component	Sig.
Level 1 predictors								
Intercept	0.181	***	1.821	n.s.	0.202	***	2.357	n.s.
Free time work day	0.961	***	0.001	n.s.	0.982	*	0.001	n.s.
Free time non-work day	0.998	n.s.	0.003	n.s.	0.998	n.s.	0.004	n.s.
Paid job					1.119	**	0.054	n.s.
Age					1.001	n.s.	0.000	n.s.
Sex					1.044	n.s.	0.067	n.s.
Cohabiting					0.960	n.s.	0.342	n.s.
Children					1.020	n.s.	0.355	n.s.
Level 2 predictors								
SCV 2002					0.931	n.s.		
SCV 2005					0.927	n.s.		
N level 1	3552				3552			
N level 2	297				179			

Note: $*p \leq 0.05$, $**p \leq 0.01$, $***p \leq 0.001$; n.s. = not significant.

such as employment status and having children. The same is found for claims of temporary busyness ("come back at another time").

These results suggest that when sample units claim they are too busy or have no time, or when they express to the interviewer that he/she needs to come back at another time, it can be a genuine signal of busyness that needs to be taken into account in order to try to find a more suitable moment for participation in data collections. It also indicates that for these "converted" initial negative participators with busyness claims in the SCV surveys, the Newtonian hypothesis could be the most fitting: although they seem to be genuinely busier, these busy sample units still somehow find the time to participate regardless if a more convenient moment is found. As for the statements of temporary busyness, we also found an effect of lack of time on week/workdays. The latter also points to chronic busyness but does not really allow us to determine whether there also was a temporary moment of extra busyness when the specific reaction to come back at another time as response to the survey request was uttered.