Journal of Official Statistics
vol. 31, i. 1 (2015)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 1–30, http://dx.doi.org/10.1515/JOS-2015-0001

Face-to-Face or Sequential Mixed-Mode Surveys Among Non-Western Minorities in the Netherlands: The Effect of Different Survey Designs on the Possibility of Nonresponse Bias

Johannes W.S. Kappelhof[1]

This article compares the quality of response samples based on a single mode CAPI survey design with the quality of response samples based on a sequential mixed-mode (CAWI-CATI-CAPI) survey design among four non-Western minority ethnic groups in the Netherlands. The quality is assessed with respect to the representativity of the response samples and the estimated potential for nonresponse bias in survey estimates based on auxiliary variables and the response rate. This article also investigates if these designs systematically enhance response rates differently among various sociodemographic subgroups based on auxiliary variables. Also, costs and cost-related issues particular to this sequential mixed-mode design are discussed. The results show that sequential mixed mode surveys among non-Western ethnic minorities in the Netherlands lead to less representative response samples and show more potential for nonresponse bias in survey estimates. Furthermore, the designs lead to systematic differences in response rates among various sociodemographic subgroups, such as older age groups. Both designs also cause some of the same sociodemographic subgroups to be systematically underrepresented among all non-Western ethnic minority groups. Finally, the results show that in this instance the cost savings did not outweigh the reduction in quality.

Key words: Survey design; sequential mixed-mode survey; nonresponse bias; non-western ethnic minorities; representativeness.

1. Introduction

In general population surveys, minority ethnic groups tend to be underrepresented (Feskens 2009; Groves and Couper 1998; Schmeets 2005; Stoop 2005). At the same time, national and international policy makers need specific information about these groups, especially on issues such as socioeconomic and cultural integration (Bijl and Verweij 2012). That is why separate surveys among the main minority ethnic groups, that is non-Western minorities, continue to be necessary in the Netherlands. However, large-scale surveys are costly, and surveys among minorities are even more expensive per completed interview than general surveys, due to the lower response rates among minorities. It is therefore of great importance to determine which strategies are effective for surveying ethnic minorities, while maintaining an acceptable level of quality and minimizing the costs.

One important part of the survey design is the data-collection mode (face-to-face, telephone, web or paper). These modes vary greatly not only in costs, but also in the

[1] The Netherlands institute for Social Research/SCP, The Hague, P.O. Box 16164, The Netherlands. Email: j.kappelhof@scp.nl
Acknowledgment: I would like to thank E.D. de Leeuw, I.A.L. Stoop and some anonymous reviewers for their very helpful comments on previous drafts of this article.

probability of completing an interview, especially among non-Western minorities (Feskens et al. 2010). There are reasons to believe that these groups may not be as well represented if a survey is conducted by means of less expensive data-collection modes as compared to a single-mode face-to-face survey. Telephone, web and mail questionnaires all lead to increased nonresponse due to higher refusal rates, a higher prevalence of functional illiteracy and/or lower penetration rates of modes compared to face-to-face (Dagevos and Schellingerhout 2003; Feskens 2009; Feskens et al. 2010; Gijsberts and Iedema 2011; Kappelhof 2010; Kemper 1998; Korte and Dagevos 2011; Schmeets 2005; Schothorst 2002; Van Ingen et al. 2007; Veenman 2002).

Despite the known limitations of other modes of data collection, there is a strong push to explore the possibility of employing less expensive methods of data collection among non-Western minorities. One possible way of reducing costs and dealing with the additional nonresponse brought about by the different modes is through the use of a sequential mixed-mode survey (De Leeuw 2005).

This article sets out to investigate:

1. how the use of a sequential mixed-mode design in surveys among non-Western minorities in the Netherlands affects the quality of the *response* sample (i.e., the composition of the group of respondents) compared to a single-mode face-to-face design, and how these two designs can potentially impact nonresponse bias. This will be referred to as the *overall quality* research question.
2. whether these designs systematically enhance response rates differently among various socio-demographic subgroups among non-Western minorities. This will be referred to as the *systematic differences* research question.
3. Finally, we will discuss costs and cost-related issues particular to this sequential mixed-mode design that are relevant in the quality versus costs trade-off decision.

The data used in this study come from a large-scale survey design experiment. Two random samples were drawn from each of the four largest non-Western minority populations living in the Netherlands. Subsequently, one sample was assigned to a face-to-face computer-assisted personal interviewing (CAPI) design and the other sample was assigned to a sequential mixed-mode design using computer-assisted web interviewing (WEB), computer-assisted telephone interviewing (CATI) and face-to-face CAPI. The fieldwork for both survey conditions was conducted simultaneously by Gfk Netherlands and lasted from November 2010 until June 2011.

In this article, we are analyzing exclusively the representativity of the *response* samples and the estimated potential for nonresponse bias based on auxiliary variables and the response rate. However, we shall not compare actual estimates of substantive variables from both survey designs as an indication of the nonresponse bias related to the estimates, given that, in this experimental design, observed differences can also be (partly) caused by mode effects in the sequential mixed-mode design (De Leeuw 2005; De Leeuw et al. 2008; Dillman and Christian 2005; Voogt and Saris 2005). Furthermore, sampling error can also contribute to observed differences, although this can be estimated.

The article presents a brief overview of the main difficulties in data collection resulting in nonresponse when surveying non-Western minorities and how survey design can reduce these difficulties. The data and methods section describes the experiment in more detail

and the methods used to answer our research aims. This is followed by the results of the analysis and the subsequent conclusion and discussion.

2. The Underrepresentation of Non-Western Minorities in Population Surveys in the Netherlands and Survey Design Choices

Statistics Netherlands uses the following official definition to describe a non-Western person in the Netherlands: "Every person residing in the Netherlands of whom one or both parents were born in Africa, Latin America, Asia (excluding Indonesia and Japan) or Turkey (Reep 2003)". A further distinction is made between first generation (born in Africa, Latin America and Asia (excluding Indonesia and Japan) or Turkey and moved to the Netherlands) and second generation (born in the Netherlands, but one or both parents were born in Africa, Latin America and Asia – excluding Indonesia and Japan – or Turkey). Indonesian and Japanese immigrants are seen as (more similar to) Western minorities based on their socioeconomic and sociocultural position, which mainly involves persons born in the former Dutch East Indies (Indonesia) and employees working for Japanese companies with their families. In 2011, non-Western minorities made up about 11% of the population in the Netherlands (CBS-Statline).

The main reason for the underrepresentation of non-Western minorities in population surveys in the Netherlands is nonresponse. A distinction can be made between direct causes and correlates for nonresponse. For instance, a direct cause would be language problems or the higher rate of illiteracy, especially among older non-Western immigrants (Feskens et al. 2010). A correlate would be that non-Western minorities more often tend to live in the larger cities in the Netherlands. Big-city dwellers in general are more difficult to contact and refuse more often (Groves and Couper 1998; Stoop 2005).

Adapting the survey design in such a way that these direct causes of nonresponse are addressed may reduce the nonresponse among non-Western minorities. Language difficulties stop being an issue if the design includes a translated questionnaire. Functional illiteracy ceases to be a problem when the interviews are conducted by interviewers who read out the questionnaire. Moreover, the use of the telephone for interviews increases the number of refusals among non-Western minorities to an incomparable degree as opposed to native Dutch or to a face-to-face mode and should therefore be avoided (Schothorst 2002).

Other cultural differences influencing nonresponse may also be reduced by specific survey design choices. For example, the use of interviewers with a common ethnic background: not only do they speak the language, but they are also aware of the proper etiquette for approaching the sampled persons. An often overlooked cause of nonresponse is the timing and length of the fieldwork. Especially among some of the ethnic minority groups, it is not uncommon to go on an extended holiday to their country of origin during the summer. Sometimes there is also a mismatch between religious holidays of ethnic groups and the way the agency plans the fieldwork (Kemper 1998; Schothorst 2002; Veenman 2002).

Sampling frame errors and especially undercoverage provide another reason why non-Western minorities are underrepresented in population surveys in the Netherlands. Undercoverage occurs when not all elements of the target population can be found in the

sampling frame (Groves 1989). In the Netherlands, (semi)-governmental and scientific institutes mainly use the postal data service (delivery sequence file) or population register as a sampling frame. Both frames suffer from frame errors, such as mobility of the sample units, no known address of the sample units, slow registration of the sample units or death of the sample units. Some of these causes occur far more often among non-Western minorities, such as mobility or no known address of sample units (Feskens 2009; Kappelhof 2010).

3. Data and Methods

3.1. Data

The Dutch Survey on the Integration of Minorities (SIM) sets out to measure the socioeconomic position of non-Western minorities as well as their sociocultural integration. This survey is a nationwide, cross-sectional survey conducted every four years starting in 2006. A large-scale survey design experiment was conducted in the 2010–2011 SIM round.

In total, Statistics Netherlands drew ten samples: two random samples of named individuals were drawn from each of five mutually exclusive population strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean (including Aruba) descent and the remainder of the population (mostly native Dutch) living in the Netherlands, aged 15 years and above. The present study focuses on how different designs affect the quality of the *response* sample and how they can potentially impact nonresponse bias in surveys conducted among non-Western minorities in the Netherlands. This is why the samples containing native Dutch are excluded from this article. The analysis is therefore based on eight samples.

Based on the official definition of non-Western minorities we will use a more narrow definition to define Dutch of Turkish, Moroccan, Surinamese, and Antillean descent to include persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles or have at least one parent who was born there. In cases where the father and mother were born in different countries, the mother's country of birth is dominant, unless the mother was born in the Netherlands, in which case the father's country of birth is dominant. These four ethnic groups make up about two thirds of the total non-Western population in the Netherlands (CBS-Statline). For the purpose of brevity, they will be referred to as Turkish, Moroccans, Surinamese and Antilleans in the remainder of this article.

From each ethnic group, one sample was allocated to a single-mode face-to-face CAPI design (SM) and one sample was allocated to a sequential mixed-mode design (MM). In the SM design, a minimum of three face-to-face contact attempts had to be conducted. The SM also included a limited reissue in which unsuccessful addresses were reissued to another CAPI interviewer who had to conduct another minimum of three face-to-face contact attempts.

In the MM design, all sample units were first sent an invitation to participate via WEB. Up to two reminders were sent to nonresponding sample units. Subsequently the remaining nonrespondents with a known fixed phone number were approached using CATI. Nonrespondents were called on at least four different days in the week, at different

time periods during the day. If there was no answer or a busy signal, the number would be called more than once within the same time period. Finally, both the WEB-nonrespondents without a known (fixed) phone number and the CATI nonrespondents were approached using face-to-face interviewers (CAPI). WEB and CATI nonresponders were contacted at least three times by a face-to-face interviewer on different days and at different time periods. CATI was added as a mode, despite previous research indicating that this was not an optimal mode for surveying ethnic minorities. This was done in order to see whether this result was still valid a decade later, especially since the second-generation immigrants are much more familiar with telephones nowadays, but mostly to see if the use of CATI could potentially lead to cost savings.

In both survey designs standard response-enhancing measures were applied, such as advance letters, incentives and the possibility for potential respondents to call a toll-free number in case of questions or in order to reschedule an appointment for an interview.

This experiment used the population register as a sampling frame and the same stratified two-stage probability sampling design in all four population strata to draw the samples. In the first stage municipalities were selected proportional to size and in the second stage a fixed number of named individuals were selected. The strata variable used was municipality size and consisted of three strata: the four largest municipalities, all with a population of over 250,000; midsize municipalities with a population of between 50,000 and 250,000; and small municipalities with a population of less than 50,000. For each target group, the sample size was proportionally allocated across different municipality size strata (Table 1).

Process data and auxiliary information, also known as paradata, are potentially useful for increasing participation, for nonresponse adjustment or for evaluating potential nonresponse bias in survey estimates (Couper 2005; Kreuter 2013; Maitland et al. 2009). In this study we use the SIM fieldwork data files. These contain both process data, such as number, time, date, and outcome of contact attempt, and auxiliary information from the sampling frame about each sample unit, such as ethnicity, age, gender, first- or second-generation immigrants, municipality, and so on.

Differences Between Survey Designs

Besides the differences in administered mode and the use of a reissue phase, there is another important aspect that varied between both survey designs that could influence the results. The average length of the questionnaire differed between modes. The estimated average length of the questionnaire in the CAPI mode, based on CAPI timers, was about

Table 1. Gross sample sizes per ethnic group and design across municipality strata

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
Large municipalities	554	344	812	502	1020	633	695	429
Midsize municipalities	727	459	674	422	662	424	945	594
Small municipalities	284	176	254	162	248	150	334	210
Total	1,565	979	1,740	1,086	1,930	1,207	1,974	1,233

45 minutes. A 45-minute questionnaire was considered too long for both CATI and WEB by fieldwork experts and experts on minority research (Feskens et al. 2010). As a result, the questionnaire length for WEB and CATI has been reduced to an estimated 30 minutes.

Another difference between the designs is the value of the conditional or promised nonmonetary incentive. The use of incentives has a proven positive effect on response rates (Dillman 2007; Groves and Couper 1998; Singer et al. 1999; Singer et al. 2000; Singer 2002). In both designs a gift certificate was used as a promised incentive. In the SM design these gift certificates were worth €10. In the MM design the amount varied: €7.50 in the WEB mode and €10 in the other modes. As mentioned above, a maximum of two reminders was sent during the WEB phase to nonresponding sampled persons. After the second reminder the worth of the conditional non-monetary incentive was increased to €12.50. As both designs used conditional incentives and the difference in value was rather small, we believe this difference between survey conditions to have a minor impact on the results.

Differences in Survey Design Between Ethnic Groups

A recent survey conducted by Statistics Netherlands among the four largest non-Western minorities discovered that approximately 14% of the sample were nonrespondents due to language problems (Feskens 2009). Results from other surveys among the same minorities groups in the Netherlands showed that nonrespondents who are not able to read or speak Dutch are found mostly among the Turkish and Moroccan populations (Kappelhof 2010). For the SIM survey, auxiliary information about ethnicity, age, gender, municipality, and status as first- or second- generation immigrants was available in the sample frame data for all sampled persons. This allowed for a tailored approach for the sampled persons. Two types of tailoring were used in both arms of the experiment to increase response. They mainly have to do with anticipated language difficulties, but also with anticipated cultural differences. Research has shown that a greater cultural familiarity due to a shared ethnic background of interviewer and respondent may also be a factor in increasing the willingness to respond (see for instance Moorman et al. 1999).

The first type of tailoring was the use of translated questionnaires and advance letters. These were used in both designs in all modes (WEB, CATI, and CAPI), but only among the Moroccan and Turkish samples. Furthermore, a phonetically translated Berber version was available as an aid for the interviewer. This is a spoken (i.e., not written) language that many Moroccans living in the Netherlands have as their mother tongue. The answers were filled in the CAPI program in either Dutch or Moroccan Arabic. There was no need to translate questionnaires or advance letters for Surinamese or Antilleans. Dutch is the mother tongue for many, if not all persons of Surinamese or Antillean origin.

The second type of tailoring is the assignment of sample units to an interviewer with a shared ethnic background. In each design, all sampled persons of Moroccan or Turkish origin were contacted by a *bilingual* interviewer with a shared ethnic background during the face-to-face (and telephone) phase. In both the single- and mixed-mode design, about half of the sampled persons of Surinamese or Antillean origin in the telephone and/or face-to-face phase were approached by interviewers with a shared ethnic background. The other half of each sample was approached by either Dutch interviewers or interviewers with another ethnic background. The allocation of Surinamese and Antillean sample units to

interviewers with a shared ethnic background was based on the availability of an interviewer with a shared ethnic background in the area.

3.2. Methods

A standard measure for judging the quality of a *response* sample is the response rate, despite the fact that it is not a direct measure and also a poor indicator of nonresponse bias (Biemer and Lyberg 2003; Groves and Peytcheva 2008). In the last few years, several other quality indicators have been developed that provide insight into the existence of nonresponse bias in survey estimates requiring somewhat weaker assumptions, such as *missing at random* (MAR) (Särndal 2011; Särndal and Lundström 2010; Schouten et al. 2009; Wagner 2010) or the weakest assumption, *missing not at random* (MNAR) (Andridge and Little 2011), and allow us to estimate its size. In order to answer our first research question – *overall quality* – we will use, next to the response rate, two approaches to evaluate how both designs affect the quality of the *response* samples and potential nonresponse bias in survey estimates for each design. In order to answer the second research question – *systematic differences* – differences in response propensity between sociodemographic subgroups, based on sample frame variables, are analyzed.

The First Approach for Assessing the Overall Quality (R1-1)

As a first approach for assessing the overall quality of the *response* samples, the representativity or R-indicator and the estimated maximal absolute *standardized* bias are used (Schouten et al. 2009). The R-indicator is a measure that describes how well the *response* sample reflects (i.e., how representative it is of) the population of interest, based on a certain number of background variables (Schouten and Cobben 2007; Schouten and Cobben 2008; Schouten et al. 2009). Obviously, this representativity only applies to the variables included in the model for estimating this measure and the response probability depends on these observed data only. One very important prerequisite is that the R-indicator needs complete (frame) data on all sample members: respondents and nonrespondents. This might not always be available. The R-indicator evaluates the differences in the estimated average response propensities between all strata, based on the variables included in the model from the available frame data. Response is considered representative if the response propensities are constant across the sample, which corresponds to a missing completely at random mechanism (Andridge and Little 2011, 154; Little and Rubin 2002).

Schouten et al. (2009, 107) show that "the R-indicator can also be used to set upper bounds to the non-response bias and to the root mean square error (RMSE) of adjusted response means." The following equation (Eq. 1) from Bethlehem et al. (2011) shows the relation between the (estimated) average response probabilities ($\widehat{\bar{\rho}}$), the R-indicator $\widehat{R}(\widehat{\rho})$, the estimated standard deviation of the survey item $\widehat{S}(y)$, and the maximal absolute bias $\widehat{B_m}(\widehat{\rho}, y)$.

$$\widehat{B_m}(\widehat{\rho}, y) = \frac{(1 - \widehat{R}(\widehat{\rho}))\widehat{S}(y)}{2\widehat{\bar{\rho}}} \tag{1}$$

For an unambiguous comparison, Bethlehem et al. (2011) use the Cauchy-Schwarz inequality to factor out the S(y). This results in the estimated maximal absolute *standardized* bias (Eq. 2):

$$\widehat{B_m}(\widehat{\rho}, y) = \frac{\left(1 - \widehat{R}(\widehat{\rho})\right)}{2\widehat{\widehat{\rho}}} \tag{2}$$

The Second Approach for Assessing the Overall Quality (R1-2)

As a second approach for assessing the overall quality of the *response* samples the fraction of missing information estimates are used (Wagner 2008; 2010). The fraction of missing information (FMI) originates from the framework of multiple imputations (Dempster et al. 1977; Rubin 1987). It is a method used for incorporating uncertainty due to missing values in variance estimates and can be used to judge the efficiency of multiple imputations. FMI is defined as the ratio of the between-imputation variability to the total variance of the survey estimates (Wagner 2008; 2010).

The FMI is proposed as an alternative measure to the response rate to assess the quality of a sample with respect to potential nonresponse bias for a single item using all available data directly: complete case data plus paradata (sample frame data and process data) (Wagner 2008; 2010).

If the FMI is below the nonresponse rate it will serve as an alternative quality indicator to the response rate. Furthermore, provided we choose the correct model (i.e., the response probability depends only on the observed variables included in the model), it allows us to estimate the potential nonresponse bias for a specific survey item.

The $\widehat{B_m}(\widehat{\rho}, y)$ and the FMI approach differ in the way they estimate how nonresponse bias can impact the survey estimate. For instance, the $\widehat{B_m}(\widehat{\rho, y})$ presented in Equations (1) and (2) is an estimate of the upper bound nonresponse bias for a hypothetical survey item, under the scenario where nonresponse correlates maximally to this variable (Schouten et al. 2011). It is based on the auxiliary variables in the model and an assumed correlation between these variables and the hypothetical survey item. There is no item-specific estimate for nonresponse bias.

Wagner's approach is designed to estimate the effect of nonresponse bias on the actual item level. In his approach, Wagner (2010) assumes that the missingness of the variable Y is independent of Y after conditioning on the covariates included in the model. This relates to a missing at random assumption (Andridge and Little 2011). Andridge and Little (2011) even extended the approach to MNAR models.

Given the difference in survey and item level-based estimates of nonresponse bias, it is interesting to compare the results of the $\widehat{B_m}(\widehat{\rho}, y)$ with the FMI approach to see whether they yield similar results. To this end we will compare the FMI results of multiple items and compare the combined results to the outcome of the $\widehat{B_m}(\widehat{\rho}, y)$.

Assessing Systematic Differences (R2)

Sometimes certain sociodemographic subgroups, such as young males, can be expected to have a different position or opinion on important research topics, such as having a job or

the attitude on sociocultural integration. When they are under or overrepresented in the response sample, the results with respect to these research questions may be biased.

It is therefore important to see whether the different designs systematically affect the response composition of surveys among non-Western minorities and how they affect the response composition. To answer our second research question, to see whether the survey designs systematically cause different sociodemographic subgroups to be over- or underrepresented in the response samples among non-Western minority groups, partial R-indicators will be used (Schouten et al. 2011; Schouten et al. 2012; Shlomo et al. 2009).

These sociodemographic subgroups can be determined based on variables included in the model used to estimate the R-indicator. A partial R-indicator on a variable level shows the contribution of a specific background variable included in the model to the overall lack of representativity of the final sample. A partial R-indicator can also be calculated on a category level to ascertain the contribution to the lack of representative response separately for each category.

There are *unconditional* and *conditional* partial R-indicators for discrete variables and categories. The *unconditional* partial R-indicator on a variable level can be used to make comparisons between surveys (Shlomo et al. 2009, 7). It measures the variability of the response propensities between the different categories of a variable. The larger the variability, the greater the contribution to the lack of representativity. This indicator is non-negative and bounded above by 0.5 (Schouten et al. 2011, 236).

The values of the *unconditional* partial R-indicators on a category level may take values between -0.5 and 0.5 (Schouten et al. 2011, 236). A negative value indicates an underrepresented category and a positive value indicates an overrepresented category and zero (0) means representative.

The *conditional* partial R-indicator on a variable level measures the contribution of a variable to the lack of representative response, adjusted for the impact of the other variables included in the model (Schouten et al. 2011, 237). It tries to isolate the part of the nonrepresentative response that can be attributed to a specific variable. The conditional partial R-indicator on a variable level can take on any value in the interval [0, 0.5.]

The values of the *conditional* partial R-indicator on the category level range from 0 to 0.5 and show the conditional contribution of a category to the lack of representative response. The higher the value, the larger the contribution of the category to the lack of representativity.

4. Results of the Comparison of Single- and Mixed-Mode Designs Among Ethnic Minorities

4.1. Results on Overall Quality (R1-1): Representativity and the Maximal Absolute Standardized Bias

"When indicators are used to compare multiple surveys, and partial R-indicators could be part of such a comparison, then generally available auxiliary variables should be selected for which literature has shown that they relate to nonresponse in most if not all surveys" (Schouten et al. 2011, 15). In this section, the paradata used consists of the auxiliary sample frame variables *Age group*, *sex*, *municipality size* and *immigration generation*. All

these variables have shown a large variability between the categories on the propensity to respond (see for instance Feskens et al. 2010; Groves and Couper 1998; Stoop 2005). No other complete frame data was available for inclusion in the analysis. The final R-indicator model we used consisted of *Age group* (six categories: 15–24; 25–34; 35–44; 45–54; 55–64; above 64 years); *Sex* (male and female); *Municipality size* (three categories: large, middle and small) and *Immigration generation* (first and second immigration generation), plus three interaction terms: *Age group* ∗ *Municipality size*; *Immigration generation* ∗ *Sex*; and *Immigration generation* ∗ *Municipality size*.

For this study we used the AAPOR definition 1, the minimum response rate, to calculate the response rate (AAPOR 2011). Looking at the results in Table 2, the following pattern emerges. In each of the four mixed-mode samples a significantly higher response rate was achieved in comparison to their single-mode counterparts. However, the representativity of each of the single-mode *response* samples is significantly higher than each of the corresponding mixed-mode *response* samples. So, despite achieving the highest response rate, the mixed-mode *response* sample does not result in the best response composition with respect to the variables included in the model.

The $\widehat{B_m}$ takes into account both the response rate and the response composition with respect to the variables in the model (Eq. 2). The $\widehat{B_m}$ shows similar results to the R-indicator. The single-mode *response* samples all result in lower $\widehat{B_m}$ estimates than their mixed-mode counterparts.

The R-indicator shows that the SM design leads to a more representative sample compared to the MM design across and within ethnic groups, although there is no significant difference between the R-indicators of the Turkish SM and the Surinamese and Antillean MM design.

However, when the response rate is taken into account, resulting in the $\widehat{B_m}$ estimate, the SM design always leads to lower estimates for the upper bound nonresponse bias than the MM design-based estimates.

4.2. *Results on Overall Quality (R1-2): Fraction of Missing Information (FMI)*

The FMI was also used to assess how different survey designs affect the quality of the survey estimates. This was done separately for each of the four ethnic groups for both

Table 2. Response rate (RR_1), R-indicator ($\widehat{R}$), 95%-confidence interval R-indicator ($\widehat{R}^{Cl}_{0.95}$), maximal absolute standardized bias ($\widehat{B_m}$) and gross sample size (N'), separate for each ethnic group and survey design (single mode (SM) or sequential mixed mode (MM))

Ethnic group	Survey	RR_1 (%)	$\widehat{R}$ (%)	$\widehat{R}^{Cl}_{0.95}$ (%)	$\widehat{Bm}$(%)	N′
Turkish	SM	52.1	80.5*	(79.5–81.4)	18.8	1,564
	MM	54.5	76.8	(75.6–77.9)	21.4	9,78
Moroccans	SM	48.0	85.7*	(84.5–87.0)	14.8	1,737
	MM	51.7	75.8	(74.4–77.1)	23.4	1,086
Surinamese	SM	41.0	86.6*	(85.5–87.8)	16.4	1,929
	MM	43.1	80.7	(79.3–82.1)	22.4	1,203
Antilleans	SM	44.2	85.6*	(84.9–86.2)	16.4	1,973
	MM	44.4	79.1	(78.2–80.1)	23.4	1,231

Note: $*p = <0.05$. N′ based on eligible cases.

designs. To estimate the FMI the following paradata were used: the same auxiliary variables (and interaction terms) from the sample frame as for the R-indicator plus the process data variable "number of contact attempts". Dummies were used to indicate contact via Web, CATI, one face-to-face contact attempt, two face-to-face contact attempts, and so on. Web was used as the reference category.

Since the FMI is an indicator of quality at the survey variable level and we want to evaluate the quality of both survey designs, we have selected and calculated the FMI for 16 different survey items. These items cover a wide range of topics (see Appendix A). The combined results should provide us with a good indication of the overall quality of the final response sample.

We followed the guidelines provided by Graham et al. (2007) and Wagner (2008) and we used 100 multiple imputations per item to reliably estimate the FMI separately for each ethnic group within each design. Table 3 presents the summary results of the analysis and the actual FMI estimates are shown in Appendix B.

In the SM design, the majority of the items included in the analysis have an FMI below the corresponding nonresponse rate (NR). This is true among all ethnic groups. This indicates that for the majority of the survey items included in the analysis, there is less uncertainty about the (mean) values for those estimates based on the imputed data compared to the estimates based on the complete case data only.

For the MM design the reverse is true, the FMI generally being above the corresponding nonresponse rate. This tells us that, using the same model, there is more uncertainty about the imputed values based on the MM survey data, which would indicate a less balanced sample. In this case the nonresponse rate is the better indicator for the survey data quality and the potential for nonresponse bias in a survey estimate than the difference between the response sample-based estimate and the estimate based on the fully imputed dataset.

There is a clear relationship between the (non)response rate and the fraction of missing information (see for instance, Wagner 2008). The higher the response rate, the lower the expected FMI. Within each ethnic group, the SM design resulted in a lower response rate

Table 3. Summary results of the fraction of missing information estimates ($\widehat{FMI}$) and for the 16 survey items, separately per ethnic group and survey design

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
No. of items with the $\widehat{FMI}$ below NR	14	4	12	4	14	0	13	0
No. of items with the lowest $\widehat{FMI}$ when SM and MM are compared within an ethnic group	14	2	12	4	16	0	16	0
No. of items in the SM for which the $\widehat{FMI}$ is below the MM NR rate compared within an ethnic group	12		12		14		12	

Note: FMI = fraction of missing information estimate; NR = nonresponse rate; SM = single-mode survey design; MM = sequential mixed-mode survey design.

than the MM design (see for instance Table 2). We could therefore have expected that within each group the FMI estimates based on the MM design would be below the FMI estimates based on the SM design. However, when compared within an ethnic group, the FMI estimates based on the SM survey data are mostly lower than the FMI estimates based on the MM survey data. Finally, the FMI estimates based on the SM design could still be above the nonresponse rate of the MM, because many of the MM FMI estimates were above their corresponding nonresponse rate. This means that the SM FMI estimates could still be surrounded by more uncertainty than the MM estimates based on the response rate. However, the majority of the FMI estimates based on the SM design are also below the nonresponse rate of the MM design within each ethnic group (Table 3, last row). All in all, these results can be seen as an indication that the single-mode design leads to better quality estimates across the ethnic groups than the sequential mixed-mode design. However, some caution is needed because the different modes in the sequential mixed-mode design may contribute additional uncertainty about the estimates based on imputed data due to mode-related effects (a model that included type of mode was also analyzed, but yielded similar results). Furthermore, we make the assumption that our model is correct and comparable within each separate ethnic group.

Comparison of the Estimated Maximal Absolute Standardized Bias ($\widehat{B_m}$) and the Mean of the 16 Fraction of Missing Information Estimates ($\widehat{\overline{FMI}}$)

Ideally both quality indicators should produce similar results because they incorporate response rate and the sample composition information and because more or less identical models were used to estimate both sets of indicators. To this end, we have compared the eight outcomes of $\widehat{B_m}$ with the eight outcomes of the $\widehat{\overline{FMI}}$ (plus standard deviation) to check whether or not they lead to similar conclusions (Table 4). We have chosen to use the $\widehat{\overline{FMI}}$ based on all 16 survey items to obtain an overall idea about the amount of uncertainty related to imputed means based on either SM or MM survey data.

The results differ somewhat if we compare both survey designs across all ethnic groups (Table 4). For instance, the lowest $\widehat{B_m}$ does not correspond with the lowest $\widehat{\overline{FMI}}$. Also, the four lowest $\widehat{B_m}$ estimates all come from SM *response* samples, whereas this is only true for three out of the four lowest values of the $\widehat{\overline{FMI}}$. However, the results are quite similar if we compare the indicators within an ethnic group. Within each ethnic group, both $\widehat{B_m}$ and $\widehat{\overline{FMI}}$ are lower when they are based on the SM data than on the MM data. This result makes sense because, while the $\widehat{B_m}$ is designed to be comparable across surveys, the predictive value of the auxiliary variables when used directly for imputation is most likely not the same for each sample. However, it will be much more similar in the two samples from the

Table 4. The estimated maximal absolute standardized bias ($\widehat{B_m}$), the mean and standard deviation of the 16 fraction of missing information estimates ($\widehat{\overline{FMI}}$) separately for SM and MM and ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
$\widehat{\overline{FMI}}$(sd.)	44.7 (4.4)	51.0 (6.5)	50.1 (4.5)	53.3 (5.2)	54.0 (4.8)	70.2 (5.6)	49.7 (6.4)	61.4 (3.8)
$\widehat{B_m}$	18.8	21.4	14.8	23.4	16.4	22.4	16.4	23.4

same ethnic population. Still, we would gather that both estimates lead to the conclusion that the SM design outperforms the MM design.

4.3. Results on the Systematic Differences (R2): Partial R-Indicator Results

In order to answer our second research question, we want to find out whether there is a systematic impact of the survey design on the representativeness of the response across the auxiliary variable categories included in our response model. By 'systematic' we mean that the same pattern is seen across all ethnic groups. Accordingly we shall start by examining the evolution of the variation in response propensities for all variables included in the response model for the different stages of the sequential mixed-mode design, separately for each ethnic group. Next we will examine how the response samples at the different stages of the sequential mixed-mode survey compare to the response sample of the single-mode survey with respect to the variation of the response propensities.

In this section, the paradata used consists of the same four auxiliary sample frame variables. Table 5 shows the main findings of the (more or less) systematic impact that each separate mode in the sequential mixed mode had on the representativeness of the response for the variables included in our response model, separately for each ethnic group. The impact of CATI and CAPI in the sequential design shown here is conditional on the previous modes used. Also, the CATI and CAPI results refer to the unique impact and not the cumulative impact which is shown in Table 6.

Tables 5 and 6 also contain the main findings of the single-mode survey design, separately for each ethnic group. Appendix C contains the tables with the actual values of the unconditional and conditional partial R-indicators of these four variables. These tables contain the values of both the variable and category-level indicators of the various stages of the sequential mixed-mode *response* samples and the single-mode CAPI *response* samples, separately for each ethnic group.

For ease of interpretation the different stages of the sequential mixed-mode design are presented first, followed by the single-mode design (SM), separately for each group. Rows indicated with "++++" mean a consistent pattern of overrepresentation across ethnic groups of the sociodemographic category within a certain survey mode. Rows indicated with "−−−−" mean a consistent pattern of underrepresentation across ethnic groups of the sociodemographic category within a certain survey mode. Rows indicated with a combination of "+" and "0" (e.g., ++ 0 0) mean a mostly consistent pattern of representative to over representative response across ethnic groups of the socio-demographic category within a certain survey mode. Rows indicated with a combination of "−" and "0" (e.g., −− 0 0) mean a mostly consistent pattern of underrepresentative to representative response across ethnic groups of the sociodemographic category within a certain survey mode. Finally, empty rows indicate that no consistent pattern can be discerned across ethnic groups of the sociodemographic category within a certain survey mode.

The Introduction of WEB (M_{web})

The use of WEB causes differing levels of representativeness with respect to the variables included in the response model across the four ethnic groups. *Age group* and *immigration*

Table 5. Systematic impact of each separate stage in the sequential mixed-mode design and the single-mode design on the representative response of the variables included in the response model, separately for each ethnic group

	M_{web}				M_{tel}				M_{f2f}				SM			
	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A
Age group																
15-24	+	+	+	+					+	+	+	+	+	+	+	+
25-34					−	−	−	−	−	−	−	−	−	−	−	−
35-44																
45-54	−	−	0	−									+	+	+	+
55-64	−	−	−	0	0	+	+	+					+	0	0	0
>64	−	−	0	−	+	0	+	+	+	+	+	+	+	0	+	0
Gender																
Male																
Female					+	+	+	+								
Municipality size																
Large	0	−	−	−									−	0	−	−
Midsize													+	+	+	+
Small	0	0	0	+	+	+	+	0								
Immigration generation																
1st generation	−	−	−	−	0	0	+	+	+	0	+	+				
2nd generation	+	+	+	+												

Note: M_{web} = result of the introduction of WEB; M_{tel} = result of the introduction of CATI in the mixed-mode sequence; M_{f2f} = result of the introduction of CAPI in the mixed-mode sequence; S = result of the single mode; T = Turkish; M = Moroccans; S = Surinamese; A = Antilleans; '+' = overrepresented; '-' = underrepresented '0' = representative. +, 0 and − are based on whether or not zero is included in the approximated confidence interval.

Table 6. Overview of the systematic impact the different stages of the sequential mixed-mode design have on the variation in the response propensities of the variables included in the model compared to the single-mode design, separately for each ethnic group

	MM WEB vs. SM								MM WEB + CATI vs. SM								MM vs. SM							
	MM WEB				SM				MM WEB + CATI				SM				MM				SM			
	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A
Age group																								
15-24	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
25-34					−	−	−	−					−	−	−	−	−	−	−	−	−	−	−	−
35-44																								
45-54	−	−	0	−	+	+	+	+	−	−	0	−	+	+	+	+					+	+	+	+
55-64	−	−	−	0	+	0	0	0					+	0	0	0	0	−	−	0	+	0	0	0
>64	−	−	0	−	+	0	+	0					+	0	+	0					+	0	+	0
Gender																								
Male									−	−	−	−					−	−	−	−				
Female									+	+	+	+					+	+	+	+				
Municipality size																								
Large	0	−	−	−	−	0	−	−	−	−	−	−	−	0	−	−	−	−	−	−	−	0	−	−
Midsize					+	+	+	+	+	+	+	0	+	+	+	+	+	+	+	+	+	+	+	+
Small	0	0	0	+					+	+	+	+												
Immigration generation																								
1st generation	−	−	−	−					−	−	−	−					−	−	−	−				
2nd generation	+	+	+	+					+	+	+	+					+	+	+	+				

Note: T = Turkish; M = Moroccans; S = Surinamese; A = Antilleans; '+' = overrepresented; '-' = underrepresented; '0' = representative. +, 0 and − are based on whether or not zero is included in the approximated confidence interval.

generation show a strong collinear response behavior among the Turkish and the Moroccans (see unconditional and conditional partial R-indicators in Appendix C). This was to be expected, since Turkish and Moroccan immigration only started in the mid-1960s and therefore second-generation immigrants over the age of 45 hardly exist (CBS-Statline). The first-generation immigrants were mostly men who came to the Netherlands for work. Partner reunification only started in the mid-seventies. Our data suggest that across all ethnic groups the young (15–24) and second-generation sampled persons find it easier to respond via WEB. The older (45 upwards) and first-generation sampled persons seem to be systematically underrepresented. Furthermore, there is also a systematic effect of WEB across the ethnic groups when it comes to *municipality size*. Persons from large cities are less inclined to participate via WEB. Finally, the use of WEB does not appear to have a systematic impact on *gender* across the ethnic groups.

The Introduction of CATI in the Sequence (M_{tel})

The success of the CATI mode was quite limited, resulting only in a very modest increase in response across the ethnic groups. Therefore the introduction of CATI in this sequence had a limited impact on the representativeness of response for the variables included in the response model. However, CATI does attract a very selective response group. The use of CATI in this sequence mainly results in female respondents, older respondents, first-generation respondents and respondents who live in small municipalities.

The Introduction of CAPI in the Sequence (M_{f2f})

The introduction of CAPI as the final mode of contact in the sequential mixed-mode design has a systematic effect on *age group* and *immigration generation* across the ethnic groups compared to WEB+CATI. With respect to *age group,* the face-to-face interviewers get either young (15 to 24) and/or older (above 64) persons to respond, but fail to get persons in the age of 25 to 34 to respond. Finally, face-to-face interviewers are able to get first generation immigrants to respond across all ethnic groups. Interestingly enough, there seems to be no systematic effect for *gender* or *municipality size* when CAPI is introduced as the final mode in this sequence.

SM: the Use of CAPI Only

The use of CAPI as a single mode of surveying ethnic minorities has a strong impact on the way different age categories are represented in the response. Persons aged 25 to 34 do not respond well and are underrepresented across all ethnic groups. The SM design also systematically results in an overrepresentation of persons aged 15 to 24. With respect to the upper three age categories, the SM design also causes these categories to be somewhat overrepresented, rather than representative response or an underrepresentation across all ethnic groups.

The SM design results in a systematic overrepresentation of persons living in midsize cities. It also leads to an underrepresentation of persons living in large cities, although among Moroccans the response is more or less representative. Finally, the SM design did not seem to have a systematic effect on *gender* or *immigration generation* across the different ethnic groups.

Partial R-Indicator Comparison Between the Different Survey Designs

The partial R-indicators on the variable level show some significant differences in the variation of the response propensities for the variables included in the response model (see Appendix C). This means that the use of different survey designs (or intermediate mode combinations of the MM design) causes different response compositions and that the size of the variation in response propensities is dependent on ethnic group, mode and variable. For instance, the use of WEB does not lead to a larger variation of the response propensities than the SM design for all the variables included in the response model, but it is dependent on the interaction between the response variable and ethnic group.

The differences in the variation of response propensities between different survey designs can also be the result of the same sociodemographic categories being more heavily under or overrepresented. For example, both the WEB and SM samples result in an overrepresentation of persons aged 15 to 24, but they differ in the degree of overrepresentation.

In order to gain a better understanding of the advantages and disadvantages of (combinations of) the current sequential mixed-mode survey design compared to a single-mode CAPI survey design, the results of the former are compared to the results of the latter in a more detailed manner.

For this comparison we will focus on whether the different survey designs cause the same or different sociodemographic categories to be systematically over- or under-represented across ethnic groups or whether this is dependent on ethnic group.

MM WEB Versus SM

The first step of the MM design (WEB only) and SM design causes some of the same categories to be under- or overrepresented (Table 6). For instance, both result in an overrepresentation of persons aged 15 to 24. Secondly, both mostly result in a small to rather large underrepresentation of big city dwellers and a representative response or overrepresentation of persons from midsize municipalities.

WEB only and the SM design also lead to the systematic under- or overrepresentation of different categories across all ethnic groups. The use of WEB usually results in an underrepresentation of the upper age categories, whereas the use of the SM design more often results in an overrepresentation of the upper age categories. Furthermore, the SM design systematically leads to an underrepresentation of persons aged 25 to 34, whereas for WEB this depends on the ethnic group. Furthermore, the use of WEB leads to a systematic underrepresentation of first-generation immigrants, which is not the case in the SM design.

An interesting result is the absence of a systematic impact of WEB only and the SM design for *gender* across the ethnic groups. As it turns out, both WEB only and the SM design lead to an over- or an underrepresentation of males (or females), dependent on ethnic group.

MM WEB+CATI Versus SM

The use of CATI as a second step in the mixed-mode sequence resulted in a low response and is therefore not recommended for ethnic minority groups. As a result of the low

response rate, the impact on the response composition is rather small and marked by the same differences and similarities found in the WEB versus SM comparison. However, because of the very selective response group in CATI, the systematic differences between WEB+CATI and the SM design have decreased somewhat for the upper age categories. Furthermore, the WEB+CATI design leads to a systematic underrepresentation of men and systematic overrepresentation of women, as opposed to the SM design.

MM Versus SM

The samples of the complete MM design show some interesting similarities with the SM design across the ethnic minorities. Both designs lead to a systematic overrepresentation of persons aged 15 to 24 and an underrepresentation of persons aged 25 to 34. They also yield the same sort of result when it comes to *Municipality size*. They both result in a systematic underrepresentation of big city dwellers and an overrepresentation of persons from midsize municipalities.

Both designs also lead to some systematic differences with respect to sociodemographic categories. First of all, the upper age categories systematically tend to be somewhat overrepresented in SM, whereas this is not a systematic finding in the MM. The opposite is actually true for persons aged 55 to 64. There is a tendency for this age group to be underrepresented in the MM. The MM design also results in an underrepresentation of men and first-generation immigrants, as opposed to the SM design. However, the underrepresentation of first-generation immigrants in MM is less severe than in the WEB+CATI design.

4.4. The Cost Perspective

The use of a sequential mixed-mode design instead of a single-mode CAPI design has the potential to greatly reduce the costs of the survey. Theoretically, the largest cost savings are made when the sequential mixed-mode design introduces the most inexpensive mode (web or postal) first and follows up with increasingly more expensive, interviewer-assisted modes. Furthermore, this can generate economies of scale when the sample size increases.

However, there are costs and cost-related considerations which are either unique or amplified in case of a sequential mixed-mode design as compared to a single-mode CAPI design that easily can be overlooked. These are especially relevant when sample sizes are relatively small and the known survey difficulties in connection with specific target populations require the use of a CAPI mode.

First of all, there are the extra costs related to questionnaire development and interviewer training. These costs can increase because the questionnaire has to be developed to be suitable for every mode and administered in different interviewer-assisted modes. From this point of view, CATI is not very cost effective as a mode among non-Western minorities in this design: only 1.3% to 6% of the sampled persons in the different ethnic groups responded via CATI.

Secondly, information costs money and, compared to a face-to-face survey design, the use of a sequential mixed-mode design limits the amount of information that can be gathered. In this experiment, the WEB and CATI questionnaire was reduced to about

two-thirds of the length of the CAPI questionnaire. This means that the cost per survey question can actually increase in a sequential mixed-mode survey.

Thirdly, time is money: the length of the fieldwork period can increase because of the use of a sequential mixed-mode design. Each mode needs a certain amount of time to be used to its full potential. For instance, in this study the second mode (CATI) was only introduced one and a half months into the fieldwork period. The need to wait for each mode to reach its full potential was the main reason for which the reissue in the sequential mixed-mode design had to be cut short. In addition, there are logistic costs related to conducting a sequential mixed-mode survey. It needs to be monitored quite carefully if and when a nonresponding sampled person can 'move' from one mode to the next.

Fourthly, there is a potential for a relative increase in travel costs for face-to-face interviewers. From a logistic point of view, the remaining number of nonresponding sampled persons in the CAPI phase of the MM design can be inconveniently located. This can also cause a reduction in the number of contact attempts an interviewer is able to conduct in a single day. It goes without saying when an interviewer is working on several surveys at the same time, this might not pose a problem.

A fifth, mixed-mode related cost concerns interviewer motivation and effort per face-to-face interview. Table 7 shows the ratio between the number of interviews and the total number of contact attempts conducted in the CAPI mode, separately for each ethnic group and survey design.

The ratio of face-to-face contact attempts to number of interviews is substantially higher in the MM compared to the SM. For instance, among the Turkish, for each 4.5 contact attempts that were made in the SM design, there was one interview completed, whereas in the MM design, this ratio was 5.3 to 1. Furthermore, the ratio among the Turkish and the Moroccans is a lot lower than among the Surinamese and the Antilleans. This indicates that a lot more unsuccessful contact attempts took place among the Surinamese and the Antilleans. This results not only in a lower response rate, but also in more effort per interview.

Put simply, face-to-face interviews are more expensive in terms of return when they are conducted as part of a sequential design. This result is of course to be expected since the 'easy' respondents have already participated via WEB or CATI, leaving the more reluctant or hard to reach sampled persons. However, the estimated costs of a face-to-face interview are to some extent based on the number of unsuccessful contact attempts that are made for each successful contact attempt. Therefore, the increased amount of effort needed in the MM CAPI phase when comparing the costs of a CAPI interview in a single-mode survey to a CAPI interview in a mixed-mode survey should be taken into account. This result not only has a direct financial implication; it can also lead to decreased motivation among interviewers, which in turn might lead to additional costs (bonus arrangements) or an

Table 7. Ratio of face-to-face contact attempts to number of interviews conducted in the CAPI mode during the first fieldwork phase for the SM and the MM samples, separately for each ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
Ratio	4.5	5.3	3.9	5.8	10.6	13.8	10.1	12.4

extension of the fieldwork period due to interviewers dropping out due to lack of motivation.

A final cost concern is related to analysis. It should not be forgotten that a sequential mixed-mode design will cost additional analysis time in order to check and correct for potential mode effects that can distort the results.

The eventual cost savings in this experiment, generated by using the current sequential mixed mode design instead of a single-mode face-to-face design among ethnic minority groups, amounted to between 12 to 20%, depending on how one would distribute fixed costs between both designs. However, given that this design choice also resulted in less information on the population of interest, a longer fieldwork period, additional analysis time and greater uncertainty related to the survey estimates based on both quality indicators, it can be concluded that in this instance the cost savings did not outweigh the reduction in quality.

5. Conclusion and Discussion

In this article we investigated how the use of a sequential mixed-mode – WEB-CATI-CAPI –design affects the quality of the *response* sample compared to a single-mode face-to-face CAPI design in surveys among non-Western minority groups in the Netherlands, as well as how these different survey designs may impact nonresponse bias on survey estimates. Statistics Netherlands drew two random samples from each of the four largest non-Western minority populations living in the Netherlands. In each ethnic group, one sample was assigned to a sequential mixed-mode design and a one sample to single-mode face-to-face CAPI design. This resulted in eight samples for analysis.

Furthermore, we analyzed whether the different survey designs enhance response rates to different degrees among different sociodemographic subgroups based on auxiliary variables. We also discussed costs and cost-related issues particular to this sequential mixed-mode design that are relevant in the quality versus costs trade-off decision.

Besides the response rate, we used two approaches to evaluate the quality of the *response* samples and potential nonresponse bias in survey estimates for both survey designs among non-Western minorities. The first approach was the representativity indicator (R-indicator) and the maximal absolute standardized bias ($\widehat{B_m}$) proposed by Schouten et al. (2009). The second approach was the fraction of missing information (FMI) proposed by Wagner (2008).

The sequential mixed-mode design resulted in higher response rates than the single-mode CAPI design in each of the four non-Western minority groups. However, both the R-indicator and the FMI approach showed that the single-mode CAPI survey design resulted in better quality *response* samples among non-Western minorities than the sequential mixed-mode survey design. Furthermore, the result of both the $\widehat{B_m}$ and the mean FMI analyses indicated that the potential for nonresponse bias in survey estimates is higher among the final samples based on a sequential mixed-mode design.

An analysis of partial R-indicators on the variable and category level was carried out to find out whether the survey designs enhance response rates differently among different sociodemographic subgroups. Overall, the variations in response propensities are larger in the sequential mixed-mode design than in the single-mode design for the variables

included in the model, with *age group* and *municipality size* showing the largest contributions.

The partial R-indicator analysis also showed that the sequential mixed-mode design systematically resulted in an underrepresentation of men, persons aged 55 to 64 and first-generation immigrants across all ethnic groups, but this pattern was not repeated for the single-mode survey design. On the other hand, the single-mode CAPI survey resulted in an overrepresentation of persons from the upper age categories (45+) among all ethnic groups, which was not the case for the sequential mixed-mode design. Furthermore, both survey designs systematically caused an underrepresentation of persons aged 25 to 34 as well as big city dwellers and an overrepresentation of young persons (15 to 24) and respondents from middle size municipalities. This systematic impact of the different survey designs on the response composition is important to bear in mind when a strong correlation is expected between a survey topic and specific over- or underrepresented sociodemographic subgroups.

The impact of each mode in the sequential mixed-mode design on the response composition was also assessed. WEB is a good startup mode to survey ethnic minorities, but cannot be recommendable as the only mode. WEB mostly results in response from young persons and second-generation immigrants across all ethnic groups.

CATI is not very suitable as a follow-up mode for conducting a survey among ethnic minorities in the Netherlands and should be avoided. It leads to a selective and low response due to high rates of refusals and non-contact. Furthermore, penetration rates are very low across the ethnic groups, especially if CATI is used as a second mode. Only 10 to 25% of the WEB nonresponders could be matched to a known phone number (Korte and Dagevos 2011).

CAPI remains a necessary part of any survey of non-Western minorities in the Netherlands. The introduction of CAPI in the sequential mixed-mode design increases the response among young and old (> 64) persons and first generation immigrants across all ethnic groups.

The cost savings of 12 to 20% with the current mixed-mode design did not justify the decrease in *response* sample quality as indicated by the R-indicator, $\widehat{B_m}$ and FMI. This design choice not only resulted in a lower-quality *response* sample and greater uncertainty related to the survey estimates in terms of nonresponse bias, but it also resulted in additional 'costs' in terms of loss of information due to shorter questionnaires, extended fieldwork time, and extra analysis time. These and other cost-related issues, such as the costs in terms of development, effort, and support versus return for the different modes and additional monitoring should be carefully reviewed before the decision to make use of a sequential mixed-mode design. Especially for relatively small sample sizes and known survey difficulties in connection with specific target populations, these additional costs may outweigh the expected savings.

The mixed-mode results do provide insight into how to improve the quality of the sample for surveys among ethnic minorities, while possibly reducing costs. A sequential WEB+CAPI design with a complete reissue or even targeted re-issue of nonresponding sample units from underrepresented sociodemographic subgroups seems better suited to yield a high and balanced response among ethnic groups than the current sequential mixed mode design, while also reducing the length of the

fieldwork period. This is the case provided the need for information does not exceed the optimal length of a WEB questionnaire. Furthermore, this design would still be less expensive to execute than a single mode CAPI design with a complete or targeted re-issue. In the re-issue, the nonresponding sampled persons should be assigned to other interviewers. To reduce the costs even more, one could consider reducing the number of face-to-face contact attempts to three or four during the first phase of fieldwork (Kappelhof 2014).

There are also several limitations to the current study. First of all, there are assumptions that go with the quality indicators used to assess the potential for nonresponse bias on survey estimates. Both quality indicators make use of the MAR assumption which is quite a strong assumption. Furthermore, in case of the R-indicator and the related measure of maximal absolute bias, no direct nonresponse bias estimate is possible since these measures are developed to compare surveys. In the case of the quality indicator based on the FMI approach, it is possible to provide direct estimates of nonresponse bias for a survey estimate given the MAR assumption. However, these results were not provided since the possibility of increased measurement variability because of the use of different survey modes in the sequential mixed-mode survey would distort the results too much (i.e., how much of the observed difference between the estimate based on the response rate and the imputed estimate was the result of nonresponse bias and how much can be contributed to the increased measurement variability). As a result, only the FMI estimates were presented as indicators of possible nonresponse bias occurrence in survey estimates. However, even then we have to assure ourselves that the measurement errors are the same across all response rates. If not, then comparing patterns of nonresponse across two designs without looking at the measurement errors is not as useful.

Another argument against our approach for estimating the FMI is that it is not actually necessary to fit the same model (i.e., include the same variables) to obtain the FMI of each dependent variable in order to be able to compare both designs. One may need a different set of predictor variables to obtain the best prediction for each separate dependent variable. Furthermore, as Andridge and Little (2011) argue, predictors used to predict response may differ from the predictors used to predict the outcome of substantive variables. Thus, it may be worth also considering other models to estimate and compare the FMI estimates which may lead to different results. However, our results are very consistent across ethnic groups and across different variables and present a fairly convincing picture that the response to MM design is highly selective for these specific populations. Nevertheless, future research should include several competing, but plausible (i.e., include variables known to correlate with the outcome variable) models to investigate to what extent the results are robust.

Finally, an interesting extension on the current study would be to include a quality indicator that allows for a direct estimate of nonresponse bias, but for which the model used for the estimates is based on the least restrictive assumption (MNAR), such as the proxy pattern-mixture approach of Andridge and Little (2011). This would allow for even more direct information that can be used in the cost- versus quality trade-off decision concerning which survey design is best suited to survey minority ethnic populations given financial and time restrictions.

Appendixes

Appendix A. Overview of the 16 Survey Questions Used in the FMI Approach

1	Do you see yourself as <ethnic group>? (Yes: No)
2	Are you currently employed? (Yes: No)
3	Do you consider yourself to be a member of a certain religion? (Yes: No)
4	To what degree do you consider yourself to be happy? (5-point scale)
5	Do you feel more <ethnic group> or Dutch? (5-point scale)
6	Generally speaking, how would you rate your health? (5-point scale)
7	Do you or your parents rent or own the house you live in? (rent/own/other)
8	Have you been discriminated against by native Dutch? (5-point scale)
9	In the Netherlands you get offered all the opportunities (5-point scale)
10	Do you have children? (Yes/No)
11	How satisfied are you with the Dutch society? (10-point scale)
12	How often did you visit a MD for yourself in the last two months? (0 to 60)
13	Do you own or have access to a computer to use for internet? (Yes/No)
14	It is better if the man is responsible for the finances (5-point scale)
15	How often do you experience difficulties when you have to talk in Dutch? (do not speak Dutch, often, sometimes or never)
16	How often did you do sports in the last 12 months?

Appendix B. Fraction of Missing Information Estimates (FMI in %) and the Nonresponse Rate (NR in %) for the 16 Survey Items, Separately for Each Ethnic Group and Survey Design (SM and MM)

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
$FMI_{Ethnic\ self}$	44.5	51.4	46.0	51.6	51.2	69.9	44.2	60.1
$FMI_{Employment}$	43.9	41.2	48.2	48.3	49.8	66.0	42.9	56.9
$FMI_{Religious}$	43.8	52.4	48.2	45.6	54.4	68.5	41.0	60.2
$FMI_{Happiness}$	50.7	56.1	51.3	58.1	56.7	75.0	47.3	65.6
$FMI_{Self\text{-}identification}$	53.9	63.0	58.0	56.0	56.9	74.2	56.4	57.0
FMI_{Health}	41.7	53.2	48.4	55.5	55.8	74.1	47.8	65.3
FMI_{House}	45.8	49.6	53.4	50.5	53.0	65.4	47.9	57.5
$FMI_{Discrimination\ self}$	45.3	51.2	51.7	55.9	50.9	74.7	61.5	63.5
$FMI_{Opportunities}$	47.1	56.8	55.7	57.7	56.2	72.6	60.3	61.6
$FMI_{Children}$	36.7	40.9	44.3	43.3	45.1	59.6	42.9	57.8
$FMI_{Satisfaction\_Society}$	47.5	59.5	54.5	57.6	61.4	77.1	57.4	70.8
FMI_{MD}	44.6	52.0	52.6	58.3	64.2	70.2	47.8	62.4
$FMI_{Internet}$	42.7	44.1	48.1	52.3	48.0	70.6	50.5	57.5
$FMI_{Man\_finance}$	45.1	52.6	52.4	59.8	55.7	77.0	53.5	62.6
FMI_{Speak_Ductch}	36.2	51.1	40.1	56.1	51.4	58.5	48.4	61.8
$FMI_{Sports\_frequency}$	45.3	40.6	48.1	46.4	53.8	69.1	45.3	61.1
NR	48.0	45.5	52.0	48.3	59.1	56.9	55.8	55.6
$NR_{Self\_identication^1}$	48.0	46.0	52.5	49.3	59.8	57.9	56.7	56.2
NR_{House^1}	48.4	46.3	53.3	48.8	59.1	56.9	56.0	56.1
$NR_{Discrimination\_self^1}$	48.0	46.2	53.1	49.0	59.1	57.1	56.2	56.3
$NR_{Opportunities^1}$	48.1	46.3	52.7	49.3	59.6	57.9	56.6	56.9
$NR_{Satisfied\_Society^1}$	48.2	45.7	52.6	48.6	59.2	57.6	55.9	55.8
NR_{MD^1}	49.2	47.2	54.1	51.6	59.3	58.6	55.9	57.0
$NR_{Man\_finance^1}$	48.1	45.6	52.6	49.3	59.1	57.4	56.1	55.9
N	1564	978	1737	1086	1929	1203	1973	1231

Note: [1] is corrected for item nonresponse.

Appendix C

Table C1. The unconditional variable- and category level partial R-indicators (multiplied by 1000) and response rate (RR_1), representativity indicator ($\widehat{R}$) and maximal absolute standardized bias ($\widehat{B_m}$) in %, separate for each ethnic group and survey design stage

	Turkish				Moroccans				Surinamese				Antilleans			
	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM
Unconditional																
Age group	60.8	53.1*	79.0	51.4[ac]	60.7	60.4*	87.7	20.0[abc]	34.0	27.9*	48.9	29.8[c]	37.8	35.4	34.8	34.8
15-24	43.9	39.5	61.6	21.1	35.4	37.9	70.0	8.5	23.0	12.0	18.2	11.4	18.5	8.5	18.5	9.3
25-34	7.0	−5.9	−44.0	−36.0	17.2	12.7	−11.0	−16.6	−12.2	−20.4	−34.4	−16.1	−6.2	−12.3	−26.8	−28.7
35-44	−9.2	2.9	−1.3	−13.1	−9.5	−13.5	−21.6	4.7	3.8	5.6	−0.8	−12.2	9.2	11.6	11.4	9.5
45-54	−23.1	−18.6	−22.3	13.8	−20.6	−17.6	−7.3	5.4	0.8	1.1	5.0	4.8	−21.3	−18.5	−3.4	14.3
55-64	−16.5	−19.5	2.8	14.9	−21.1	−12.6	−29.7	−0.5	−21.3	−8.4	−8.9	2.1	4.8	19.0	−2.0	−2.2
64+	−28.8	−22.1	−1.0	17.6	−34.2	−37.4	−35.6	−0.0	−1.7	10.6	27.6	17.9	−21.9	−13.6	−2.4	1.8
Sex	4.1*	16.0	18.3	37.6[abc]	37.3	37.5*	17.9	6.9[abc]	2.2*	18.4	27.9	4.6[bc]	27.9	33.8*	17.4	11.0[ab]
Male	2.8	−11.0	−12.7	−26.2	−26.1	−26.3	−12.5	4.8	−1.6	−13.3	−20.2	3.2	−19.9	−24.0	−12.4	−7.6
Female	−3.0	11.5	13.2	26.9	26.6	26.8	12.8	−5.0	1.5	12.7	19.2	−3.1	19.6	23.7	12.2	8.0
Municipality	4.0*	16.9	17.0	26.3[a]	18.9	17.6*	51.2	15.8[c]	40.6*	49.1*	31.5	44.9[c]	39.1	37.8*	80.7	45.6[c]
Large	−2.2	−13.1	−13.7	−4.7	−13.7	−11.7	−31.6	0.1	−25.6	−32.7	−16.5	−24.5	−21.3	−23.3	−65.0	−33.1
Medium	−0.1	10.6	8.7	16.5	11.8	5.3	10.3	8.2	31.6	36.1	25.3	7.9	−1.3	2.2	43.9	31.0
Small	3.3	13.3	5.0	−20.0	5.1	12.1	38.9	−13.5	−0.7	6.1	−8.7	36.8	32.7	29.6	19.0	−4.4
Immigration generation	52.0	44.6*	33.0	32.0[ab]	51.6	54.2*	65.4	17.5[abc]	41.8*	25.0*	14.1	1.1[abc]	63.7*	48.1*	18.2	3.4[abc]
1G	−29.7	−25.4	−18.9	18.3	−29.8	−31.3	−37.8	−9.7	−25.3	−15.1	−8.5	−0.6	−33.6	−25.3	−9.6	−1.8
2G	42.7	36.6	27.1	−26.2	42.1	44.2	53.3	14.6	33.3	19.9	11.2	0.9	54.1	40.9	15.4	2.9
RR_1	21.4	25.3	54.5	52.1	22.7	24.0	51.7	48.0	20.9	26.9	43.1	41.0	21.6	26.2	44.4	44.2
$\widehat{R}$	84.6	84.9	76.8	80.5	83.0	82.0	75.8	85.7	83.9	82.1	80.7	86.6	80.4	81.4	79.1	85.6
$\widehat{B_m}$	36.0	29.8	21.4	18.8	37.4	37.5	23.4	14.8	38.5	33.3	22.4	16.4	45.4	35.5	23.4	16.4
N[1]	978	978	978	1564	1086	1086	1086	1737	1203	1203	1203	1929	1231	1231	1231	1973

Note. A significant difference ($p = < 0.05$) between mode and subsequent mode within SIM MM is noted with an *; [a] = significant difference between MM WEB only and SIM SM within ethnic group; [b] = significant difference between MM WEB+CATI and SIM SM within ethnic group; [c] = significant difference between SIM MM and SIM SM within ethnic group. [d] = significant difference between different ethnic groups. Standard errors were approximated (not included here) using 1000 bootstrap replicates of the estimates. [1] Based on all eligible cases.

Table C2. The conditional variable- and category level partial R-indicators (multiplied by 1000), separate for each ethnic group and survey design

	Turkish				Moroccans				Surinamese				Antilleans			
	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM
Conditional																
Age group	35.6	33.9	71.8	60.5	36.2	34.0	60.6	23.8	32.7	35.8	52.2	31.1	27.8	34.0	29.6	37.6
15-24	19.4	17.8	46.7	39.7	11.3	10.6	39.7	6.2	6.5	7.2	14.1	11.6	2.3	4.9	15.4	10.5
25-34	15.5	15.6	50.2	36.3	17.8	13.1	31.4	18.3	22.1	27.5	38.4	15.6	5.4	10.2	22.4	30.6
35-44	10.4	14.5	13.0	22.9	8.3	6.3	12.5	9.8	12.6	12.6	6.0	13.9	13.3	15.5	10.7	10.3
45-54	10.2	7.5	14.0	6.5	4.5	2.1	11.8	9.2	14.0	11.0	10.5	4.9	11.3	11.8	1.9	15.6
55-64	6.6	12.0	7.7	7.8	13.6	4.1	17.0	2.9	12.8	5.2	6.9	3.0	12.2	23.0	3.3	3.2
64 +	20.0	13.5	5.9	11.8	24.3	26.5	22.8	2.1	4.0	12.9	29.2	19.0	16.9	11.1	3.6	2.4
Sex	2.6	17.5	20.2	36.5	33.9	34.9	15.6	5.3	4.0	19.7	29.0	5.2	31.5	34.9	17.6	11.9
Male	1.8	12.0	14.0	25.4	24.0	23.7	11.0	3.7	2.9	14.2	21.0	3.7	22.1	24.7	12.5	8.1
Female	1.9	12.6	14.5	26.1	23.9	23.6	11.2	3.8	2.8	13.6	20.0	3.6	22.4	24.7	12.4	8.6
Municipality	5.0	11.8	12.9	30.5	22.9	18.8	51.5	15.3	38.0	46.6	30.5	45.5	33.7	31.4	77.8	46.8
Large	1.6	9.5	10.2	8.3	15.8	13.5	34.2	1.1	23.7	30.8	15.4	25.3	17.6	19.1	62.2	32.8
Medium	3.3	6.4	7.1	20.3	16.4	9.1	15.6	7.9	29.5	34.4	24.4	9.1	5.0	4.0	43.2	32.5
Small	3.3	2.8	3.1	21.1	2.2	9.3	35.2	13.0	3.3	6.0	9.7	36.7	28.3	24.6	17.8	7.7
Immigration generation	16.5	16.0	2.0	45.0	17.0	21.0	9.2	21.6	37.2	29.9	20.6	1.4	54.1	45.1	7.3	3.1
1G	11.6	11.0	1.3	32.2	12.2	15.2	6.5	15.4	24.5	19.7	14.1	1.0	33.2	27.5	4.4	1.9
2G	11.7	11.6	1.5	31.4	11.8	14.5	6.5	15.2	27.9	22.4	15.0	1.0	42.7	35.8	5.8	2.4

6. References

AAPOR 2011. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. (7th edition). The 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 December 2013).

Andridge, R.R. and R.J. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27: 153–180.

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Bijl, R. and A. Verweij. 2012. *Measuring and Monitoring Immigrant Integration in Europe: Integration Policies and Monitoring Efforts in 17 European Countries*, 2012-8. Den Haag: SCP. Available at: http://www.scp.nl/english/Publications/Publications_by_year/Publications_2012/Measuring_and_monitoring_immigrant_integration_in_Europe (accessed June 2013).

CBS-Statline. Available at: http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37325&D1=a&D2=0&D3=0&D4=0&D5=0-4,137,152,220,237&D6=0,4,9,(l-1),l&HD=130605-0936&HDR=G2,G1,G3,T&STB=G4,G5 (accessed December 2013).

Couper, P. 2005. "Technology Trends in Survey Data Collection." *Social Science Computer Review* 23: 486–501. DOI: http://dx.doi.org/10.1177/0894439305278972.

Dagevos, J. and R. Schellingerhout. 2003. "Sociaal-culturele integratie. Contacten, cultuur en orientatie op de eigen groep." In *Rapportage minderheden*, edited by J. Dagevos, M. Gijsberts, and v. C. Praag, 317–362. [In Dutch: Socio-Cultural integration. Contacts, culture and focus on the own ethnic group]. Den Haag: SCP. Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2003/Rapportage_minderheden_2003 (accessed December 2012).

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

De Leeuw, E.D., D.A. Dillman, and J.J. Hox. 2008. "Mixed-Mode Surveys: When and Why." In *International Handbook Of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman, 299–316. New York: Taylor and Francis Group.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39: 1–38.

Dillman, D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New York: John Wiley & Sons Inc.

Dillman, D.A. and L.M. Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17: 30–52. DOI: http://dx.doi.org/10.1177/1525822X04269550.

Feskens, R.C.W. 2009. *Difficult Groups in Survey Research and the Development of Tailor-made Approach Strategies*. Den Haag/Utrecht: Statistics Netherlands and

Utrecht University. Available at: http://www.cbs.nl/NR/rdonlyres/8F317AA9-1074-4BF7-84EB-015D013DBB80/0/2009x11feskenspub.pdf (accessed October 2011).

Feskens, R.C.W., J. Kappelhof, J. Dagevos, and I.A.L Stoop. 2010. *Minderheden in de mixed-mode? Een inventarisatie van voor- en nadelen van het inzetten van verschillende dataverzamelingsmethoden onder niet-westerse migranten*. SCP-special 57. [In Dutch: Ethnic minorities in the mixed mode? An inventory of the advantages and disadvantages of employing different data collection methods among non-Western migrant] Den Haag: SCP. Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2010/Minderheden_in_de_mixed_mode (accessed December 2014).

Gijsberts, M. and J. Iedema. 2011. "Opleidingsniveau van niet-schoolgaanden en leerprestaties in het basisonderwijs." In *Jaarrapport Integratie 2011*, edited by M. Gijsberts, W. Huijnk, and J. Dagevos, 76–99. [In Dutch: Education level of persons who do not attend school and educational attainment in primary school]. Den Haag: SCP. (2012-3). Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2012/Jaarrapport_integratie_2011 (accessed December 2014).

Graham, J.W., A.E. Olchowski, and T.D. Gilreath. 2007. "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8: 206–213. DOI: http://dx.doi.org/10.1007/s11121-007-0070-9.

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

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

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly* 72: 167–189. DOI: http://dx.doi.org/10.1093/poq/nfn011.

Kappelhof, J.W.S. 2010. *Op maat gemaakt? Een evaluatie van enkele respons verbeterende maatregelen onder Nederlanders van niet-westerse afkomst.* [In Dutch: An evaluation of several response enhancing measures among Dutch of non-Western origin]. Den Haag: SCP. (SCP-special 53). Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2010/Op_maat_gemaakt (accessed December 2014).

Kappelhof, J.W.S. 2014. "The Effect of Different Survey Designs on Nonresponse in Surveys Among Non-Western Minorities in The Netherlands." *Survey Research Methods* 8: 81–98. Available at: http://www.surveymethods.org (accessed December 2014).

Kemper, F. 1998. "Gezocht: Marokkanen. Methodische problemen bij het verwerven en interviewen van allochtone respondenten." [In Dutch: Wanted: Moroccans. Methodological problems with obtaining response and interviewing respondents of foreign origin]. *Migrantenstudies* 1: 43–57.

Korte, K. and J. Dagevos. 2011. *Survey Integratie Minderheden 2011. Verantwoording van de opzet en uitvoering van een survey onder Turkse, Marokkaanse, Surinaamse en Antilliaanse Nederlanders en een autochtone vergelijkingsgroep*. [In Dutch: Survey on the Integration of Minorities 2011. Report on the design and implementation of the survey among Turkish, Moroccan, Surinamese, Antillean Dutch and a mainstream Dutch reference group]. Den Haag: SCP.

Kreuter, F. 2013. *Improving Surveys with Paradata. Analytic Uses of Process Information*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis With Missing Data*. New York: John Wiley & Sons.

Maitland, A., C. Casas-Cordero, and F. Kreuter. 2009. "An Evaluation of Nonresponse Bias Using Paradata from a Health Survey." In: Proceedings of the Section on Government Statistics: American Statistical Association, Joint Statistical Meetings, 2009. 370–378. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/proceedings/y2009/Files/303004.pdf (accessed December 2014).

Moorman, P.G., B. Newman, R.C. Millikan, C.K.J. Tse, and D.P. Sandler. 1999. "Participation rates in a Case-control Study: The Impact of Age, Race, and Race of Interviewer." *Annals of epidemiology* 9: 188–195. DOI: http://dx.doi.org/10.1016/S1047-2797(98)00057-X.

Reep, C. 2003. *Moeilijk Waarneembare Groepen. Een inventarisatie.* [In Dutch: Hard to survey populations: An inventory]. Voorburg/Heerlen: CBS. (H1568-03-s00).

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Schmeets, H. 2005. "De leefsituatie van allochtonen." In *Enquêteonderzoek onder allochtonen. Problemen en oplossingen*, edited by H. Schmeets en and R. van der Bie, 169–176. [In Dutch: The living conditions of Dutch of foreign origin]. Voorburg/Heerlen: CBS.

Schothorst, Y. 2002. "Onderzoek onder allochtonen: wat mag, wat moet en wat kan?" In: *Interviewen in de multiculturele samenleving. Problemen en oplossingen*, edited by H. Houtkoop en and J. Veenman, 101–116. [In Dutch: Survey research among Dutch of foreign origin: What may be done, what has to be done and what is possible?]. Assen: Koninklijke Van Gorcum.

Schouten, B. and F. Cobben. 2007. *R-Indexes for the Comparison of Different Fieldwork Strategies and Data Collection Modes*. Voorburg/Heerlen: CBS. (Discussion Paper 07002). Available at: http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf (accessed December 2013).

Schouten, B.F. and F. Cobben. 2008. *An Empirical Validation of R-Indicators*. Voorburg/Heerlen: CBS. (Discussion Paper 08006) Available at: http://www.risq-project.eu/papers/cobben-schouten-2008-a.pdf (accessed December 2013).

Schouten, B. F. Cobben, and J. Bethlehem. 2009. "Indicators of Representativeness of Survey Nonresponse." *Survey Methodology* 35: 101–113. Available at: http://www.risq-project.eu/papers/schouten-cobben-bethlehem-2009.pdf (accessed December 2013).

Schouten, B. J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring, and Improving Representativity of Survey Response Through R-Indicators and Partial R-Indicators." *International Statistical Review* 80: 382–399. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2012.00189.x/abstract (accessed December 2013).

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativity of Response." *Journal of Official Statistics* 27: 231–253.

Shlomo, N., C. Skinner, B. Schouten, T. Carolina, and M. Morren. 2009. *Partial Indicators for Representative Response*. Rep. No. RISQ deliverable 4. version 2. Available at: http://www.risq-project.eu/papers/RISQ-Deliverable-4-V2.pdf (accessed December 2013).

Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D. Dillman, J.L. Eltinge, and R.J.A. Little, 163–177. New York: John Wiley & Sons.

Singer, E., J. Van Hoewyk, N. Gebler, T. Raghunathan, and K. McGonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." *Journal of Official Statistics* 15: 217–230.

Singer, E., J. Van Hoewyk, and M.P. Maher. 2000. "Experiments With Incentives in Telephone Surveys." *Public Opinion Quarterly* 64: 171–188. DOI: http://dx.doi.org/10.1086/317761.

Smulders, M. 2011. *Onderzoeksverantwoording Survey Integratie Minderheden 2011* [In Dutch: research description on the Survey on the Integration of Minorities 2011]. Dongen: GFK Panel Services Benelux B.V.

Stoop, I.A.L. 2005. *The Hunt for the Last Respondent: Nonresponse in Sample Surveys*. Den Haag: SCP.

Särndal, C.-E. 2011. "Three Factors to Signal Non Response Bias With Applications to Categorical Auxiliary Variables." *International Statistical Review* 79: 233–254. DOI: http://dx.doi.org/10.1111/j.1751-5823.2011.00142.x.

Särndal, C.-E., and S. Lundström. 2010. "Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias." *Survey Methodology* 36: 131–144.

Van Ingen, E., J. De Haan, and M. Duimel. 2007. *Achterstand en afstand. Digitale vaardigheden van lager opgeleiden, ouderen, allochtonen en inactieven*. [In Dutch: Lagging behind. Digital skills of lower educated, elderly, foreign origin and inactive persons]. Den Haag: SCP. (SCP- 2007/24).

Veenman, J. 2002. "Interviewen in multicultureel Nederland." In *Interviewen in de multiculturele samenleving. Problemen en oplossingen*, edited by H. Houtkoop en and J. Veenman, 1–19. [In Dutch: Interviewing in the multi-cultural Netherlands]. Assen: Koninklijke Van Gorcum.

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

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*, Michigan: University of Michigan. Available at: http://deepblue.lib.umich.edu/bitstream/2027.42/60831/1/jameswag_1.pdf (accessed December 2014).

Wagner, J. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74: 223–243. DOI: http://dx.doi.org/10.1093/poq/nfq007.

Received March 2013
Revised December 2014
Accepted December 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 31–59, http://dx.doi.org/10.1515/JOS-2015-0002

Validating Sensitive Questions: A Comparison of Survey and Register Data

Antje Kirchner[1]

This article explores the randomized response technique (RRT) – to be specific, a symmetric forced-choice implementation – as a means of improving the quality of survey data collected on receipt of basic income support. Because the sampled persons in this study were selected from administrative records, the proportion of respondents who have received transfer payments for basic income support, and thus the proportion of respondents who should have reported receipt is known.

The article addresses two research questions: First, it assesses whether the proportion of socially undesirable responses (indication of receipt of transfer payments) can be increased by applying the RRT. Estimates obtained in the RRT condition are compared to those from direct questioning, as well as to the known true prevalence. Such administrative record data are rare in the literature on sensitive questions and provide a unique opportunity to evaluate the 'more-is-better' assumption. Second, using multivariate analyses, mechanisms contributing to response accuracy are analyzed for one of the subsamples.

The main results can be summarized as follows: reporting accuracy of welfare benefit receipt cannot be increased using this particular variant of the RRT. Further, there is only weak evidence that the RRT elicits more accurate information compared to direct questioning in specific subpopulations.

Key words: Randomized response technique; social desirability; validation data; welfare receipt; unemployment benefit II.

1. Introduction

Surveys that collect data on welfare and unemployment receipt often find that respondents underreport this kind of information. In German surveys the known extent of underreporting of receipt of basic income support, a form of means-tested social security payment called 'Unemployment Benefit II' (UB II), ranges between 9 percent and 17 percent (Kreuter et al. 2010, 2014). One potential motivation for underreporting might be the sensitive nature of the topic: by underreporting, respondents avoid interviewer disapproval, embarrassment, and answer in a socially desirable manner (Tourangeau and Yan 2007). The main question the following paper addresses is whether alternative questioning formats, such as the randomized response technique (Warner 1965), can be

[1] Institute for Employment Research (IAB), Regensburger Str.104, Nuremberg 90478, Germany. Email: antje.kirchner@iab.de

Acknowledgments: This study is part of a larger research project carried out with M. Trappmann, I. Krumpal and H. v. Hermanni. It has been published in part in my dissertation (Kirchner 2014). Data collection was supported by the Institute for Employment Research (IAB). I am also very grateful to S. Eckman, J. Korbmacher, F. Kreuter and the anonymous referees for their helpful comments and suggestions.

used to improve the response quality of data collected regarding welfare receipt in labor market surveys.

1.1. Background

While unintentional misreporting, for example due to recall error, can certainly be problematic in the reporting of social security receipt (Manzoni et al. 2010; Kreuter et al. 2014), special attention should be devoted to other causes of misreporting in interviewer-administered surveys. It can be reasonably assumed that survey respondents are more likely to conceal sensitive information in order to conform to perceived norms (Cialdini 2007). This, in turn, affects the validity of the prevalence estimates (Lee 1993): if this failure to report welfare receipt is systematically different for certain social groups, resulting parameter estimates, such as proportions, averages, as well as relationships between variables will be biased (Hausman 2001).

The level of 'threat' or 'sensitivity' of a question as perceived by the respondent can be established along three theoretical dimensions (Tourangeau and Yan 2007): intrusiveness, risk of disclosure and social desirability. Several of these apply to the receipt of basic income support: people apply for welfare benefits in Germany if they have been unemployed long-term or if they cannot sustain a living from their current job, that is, when the resulting income is below a legally defined threshold. Individuals receiving basic income support may not wish to report this information in a survey. Admitting to an interviewer that they either have not been able to find a job over a longer period, that they live in poverty or that they do not earn enough to support their families might be perceived as too embarrassing. The concept of 'injunctive social norms' (Cialdini 2007) – one's perception or expectation of what most others approve or disapprove of – plays a vital role in this context. Negative beliefs and prejudice about welfare recipients in the United States and Great Britain comprise anything from not being motivated enough to find a job, being uninterested in self-improvement and dishonesty, to laziness and dependence (Bullock 2006). The receipt of basic income support in Germany is associated with similar prejudice. It is thus considered socially undesirable in terms of the commonly perceived norm and negatively stigmatizing, causing embarrassment when admitting to such.

To avoid errors from (item) nonresponse and misreporting ('under-' as well as 'overreporting') due to the sensitive nature of a question, survey methodologists have suggested a range of guidelines with respect to the design of a questionnaire (for an extensive overview, see Lee 1993; Bradburn et al. 2004; Tourangeau and Yan 2007). Indirect surveying techniques, such as the randomized response technique (RRT), are strategies to reduce underreporting (Lensvelt-Mulders et al. 2005). The RRT method was originally developed by Warner in 1965 to reduce response bias arising from privacy concerns. Ever since its first implementation, the RRT has been refined in many different variants (Horvitz et al. 1967; Greenberg et al. 1969; Boruch 1971; Greenberg et al. 1971; Moors 1971; Kuk 1990; Mangat and Singh 1990; Mangat 1994). Warner's original design, the so-called unrelated question techniques, forced-response designs, Moor's design, as well as Kuk's or Mangat's variants are probably among the best-known RRT designs (for an overview of different RRT designs, estimators and applications, see Fox and Tracy 1986; Umesh and Peterson 1991; Lensvelt-Mulders et al. 2005; Lensvelt-Mulders

et al. 2005b, or Tourangeau and Yan 2007). More recent developments also account for the fact that respondents might still underreport sensitive attributes in the RRT and allow for an estimation of a so-called 'cheating' parameter (Clark and Desharnais 1998; Böckenholt et al. 2009; Van den Hout et al. 2010; Ostapczuk et al. 2011; De Jong et al. 2012).

1.2. The General Idea of the RRT

The main idea, common to all RRT variants, is to conceal a respondent's answer by using a randomizing device (e.g., coins, cards, dice, spinner), the outcome of which is only known to the respondent (Fox and Tracy 1986). In its original implementation (Warner 1965), survey respondents are – depending on the outcome of the randomizing device – directed to answer one of two logically opposing statements, such as: 'I am a recipient of unemployment benefits II,' or 'I am not a recipient of unemployment benefits II.' Respondents only answer 'Yes' or 'No' without revealing which statement they were directed to reply to. Due to this chance element, neither the interviewer nor the researcher can infer anything regarding the respondent's true status from the response given. Since the randomization mechanism – and thus the probability distribution of the misclassification – is known to the researcher, estimation of the population prevalence of the sensitive characteristic under study is possible (Fox and Tracy 1986), as are regression analyses analyzing randomized response dependent variables (Maddala 1983, 54ff.; for an overview of estimators, see Tourangeau and Yan 2007). Granted that respondents understand and trust the method, the RRT should then increase reporting accuracy and reduce measurement error resulting from social desirability concerns.

Lensvelt-Mulders et al. (2005) distinguish two main types of studies in order to assess the performance of the RRT compared to that of other techniques: comparative and validation studies. The first type of study is most commonly found when evaluating the RRT. It compares estimates derived by means of RRT to those obtained by means of standard direct questioning. The RRT is – or more generally indirect techniques are – then assumed to outperform direct questioning if it elicits higher prevalence estimates for questions that are assumed to be subject to underreporting. Researchers generally refer to this as the 'more-is-better' assumption (for an overview of studies relying on this assumption, see Umesh and Peterson 1991; Lensvelt-Mulders et al. 2005; Tourangeau and Yan 2007). These studies often use a split-ballot design, randomly assigning participants of a given survey to either direct questioning or RRT. From a validation perspective, studies relying on the more-is-better assumption provide the weakest form of validation (Moshagen et al. 2014). Alternatively, estimates from other sources in which the prevalence of the sensitive trait is known only for the population, or parts thereof, but not for the sample, can be used as a benchmark for comparison (Moshagen et al. 2014). The authors refer to this as an intermediate form of validation and point out that potential differences might be confounded with sampling bias.

In some rare instances, researchers have access to additional, auxiliary information on the subjects of investigation for evaluation of the RRT performance (for an overview, see Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013). These studies are henceforth referred to as validation studies. Validation studies provide a stronger form of

performance assessment compared to comparative studies (Lensvelt-Mulders et al. 2005). In general, two types of validation studies can be distinguished: those validating responses at the individual level and those validating responses, or rather estimates, at the aggregate level (for the same sample).

While the most powerful validation of a survey response can be achieved if a 'gold standard' or the 'true' response of a respondent is available at the individual level (Groves 1989), often this information is impossible to acquire, too costly or (legally) not accessible. However, if individual-level validation data is available, it provides a valuable resource for analyzing individual motivations that contribute to misreporting which otherwise would not be possible. The second, somewhat weaker form of validation compares RRT survey estimates to aggregate data. This information might be data that is available for certain population segments of the sample using records (such as criminal statistics) or information that is available on the sampling frame.

Many empirical studies have evaluated whether the RRT method is in fact better at eliciting reports of sensitive behavior than the direct questioning methods. In the most recent meta-analysis (Lensvelt-Mulders et al. 2005), a total of six individual-level RRT validation studies as opposed to 32 comparative RRT studies were investigated. In general, the RRT still produced some response error, albeit lower than a comparable standard face-to-face questioning: for the validation studies under investigation, in the RRT condition the mean response was underreported by 38 percent, while in the traditional face-to-face condition mean underreporting was 42 percent. One of these validation studies, conducted by van der Heijden and colleagues (2000; see also Lensvelt-Mulders et al. 2006), tested two different implementations of the RRT, a forced-response implementation and Kuk's method, against standard face-to-face questioning. Results suggest that both RRT versions yield significantly lower response error with respect to social security fraud. Other experimental studies without validation data (comparative studies based on the more-is-better assumption) also showed that the RRT increased the validity of the estimates by eliciting more truthful responses (e.g., Weissman et al. 1986; Lara et al. 2004; Lara et al. 2006).

In general, the RRT seems to elicit more honest answers and reduce social desirability bias, especially when dealing with more sensitive questions (Fidler and Kleinknecht 1977; Landsheer et al. 1999; Lensvelt-Mulders et al. 2005). For example, the pioneering validation study by Locander and colleagues (1976) relying on individual-level validation data for some items shows, that the response error for RRT is (significantly) lower compared to that of direct questioning in three out of five instances (voter registration, bankruptcy involvement, and drunken driving). While the trend – that is, the RRT eliciting higher prevalence estimates – is as expected in most validation studies on topics such as failing course grades, arrests per person or criminal convictions, some validation studies also find no significant difference between RRT and direct questioning or contrary evidence (Locander et al. 1976; Lamb and Stem 1978; Tracy and Fox 1981; Wolter and Preisendörfer 2013). More recently, other comparative experimental studies have been published questioning the validity of RRT estimates (Umesh and Peterson 1991; Holbrook and Krosnick 2010; Coutts and Jann 2011; Coutts et al. 2011; Höglinger et al. 2014). Those studies show that the RRT does not provide more valid prevalence estimates compared to direct questioning, and that the RRT provides impossible, out-of-range

estimates (Holbrook and Krosnick 2010; Höglinger et al. 2014), suggesting noncompliance with RRT instructions.

1.3. Research Objectives

The following article presents one of the few large-scale RRT validation studies using administrative record data. More precisely, it explores whether the RRT is successful in eliciting higher-quality responses regarding the receipt of basic income support. Drawing on survey data collected in a nationwide telephone survey in Germany in 2010, respondents were randomly assigned to one of two techniques: either randomized response technique or traditional direct questioning. Using administrative record data, the true percentage of respondents who have received transfer payments for basic income support and thus the percentage who should have reported receipt is known. This allows a validation of the reported percentage against the known true rate for the responding cases, hence assessing the bias of the estimates. Such administrative record data is quite rare in the literature on sensitive questions (Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013).

The study contributes to the existing RRT research and response bias in several ways: to the best of the author's knowledge, the performance of the RRT in a telephone survey has not been validated against external data (especially not with respect to the receipt of basic income support). All existing RRT validation studies implemented the RRT method in a face-to-face mode (comparing the technique with face-to-face and other modes) but never in a pure telephone setting (cf. also Lensvelt-Mulders et al. 2005; Wolter and Preisendörfer 2013). The choice of a telephone mode, however, might be perceived as more private by respondents, thus leading to more honest answers due to greater perceived social distance (Holbrook et al. 2003). While collecting data by means of the RRT has many advantages, RRT procedures also suffer from considerable disadvantages compared to direct questioning: for one, a larger sample size is needed to achieve the same statistical power (Warner 1965); second, interview duration increases due to an explanation of the application of the procedure; while third, the cognitive burden placed on respondents is higher. Validating the functioning of a telephone implementation of the RRT might prove useful, given that it is more cost efficient compared to face-to-face surveys. The study thus follows the recommendation by Lamb and Stem (1978, 617) that "each time the [RRT] method is changed or used in a different setting, further evaluation is appropriate." Furthermore, this article contributes to the literature by investigating which individual-level factors influence accurate reporting and whether these mechanisms differ across experimental conditions.

To summarize, this article addresses two research questions:

1. Can item-specific response bias in interviewer-administered telephone surveys be reduced when using the randomized response technique? This is achieved by comparing the RRT estimates with a) the true value from the administrative data and b) direct questioning (DQ) obtained from the survey data.
2. Which covariates influence response error and can the RRT contribute to diminishing response error due to perceived sensitivity?

The remainder of this article is organized as follows: Section 2 describes the experimental design, the available data, as well as the method of analysis. The results of

the experiment are found in Section 3 and the conclusions and limitations of the study in Section 4.

2. Data and Methods

The nationwide telephone survey was commissioned by the Institute for Employment Research (IAB), the research institute of the Federal Employment Agency (FEA), and was carried out by the ForschungsWerk institute from October to December 2010. This study was approved by an internal review board as part of a study investigating undeclared work (Kirchner et al. 2013). The main focus on undeclared work had design implications regarding the choice of the misclassification probabilities for the RRT design in the current study on welfare benefit receipt (see below). Due to the particular sampling design, in addition to these survey data, supplementary information is available on the sampling frame (administrative records). The combination of both data sources allows addressing the research questions stated above. The next section provides an overview of the survey data, the administrative data, the combined data, and lays out the methods of analysis.

2.1. The Survey Data

2.1.1. Sampling and Data Collection

The survey is a dual-frame survey, using two sampling frames that are maintained by the FEA. These frames consist of all registered unemployment benefit (II) recipients as well as all employed persons.

The first random sample was drawn from the FEA registers of basic income support recipients (IAB Unemployment Benefit II History (LHG) V6.03.01 and (XLHG) V01.06.00-201007). It consists of people aged 18 to 64 who were known to have received basic income support in June 2010 (henceforth referred to as UB II or benefit recipients sample). The second random sample was drawn from the register of employees that is maintained by the FEA (IAB Employment Histories (BeH) V08.04.00, Nuremberg 2010). It consists of people aged 18 to 70 who were employed in December 2009 (henceforth referred to as employee sample). For both samples the latest available registers were used.

The registers contain telephone numbers for many of the sampled individuals. Whenever there was no information available on either of the frames, an extensive telephone number research was conducted, resulting in 91.7 percent (UB II sample) and 68.2 percent (employee sample) coverage. All individuals selected into the sample received a personalized advance letter announcing the survey. During fieldwork, some of the telephone numbers turned out to be invalid. This resulted in effectively 75.8 percent (UB II sample) and 53.5 percent (employee sample) cases with working numbers. Of those cases approximately 26 percent agreed to participate in the survey. Overall 3,211 interviews were completed (UB II: 18.8 percent and employees: 16.3 percent RR1, AAPOR 2011).

2.1.2. Experimental Design and Measurement of the Dependent Variable

Individuals who were initially selected into the sample were randomized in advance into two experimental groups. To achieve approximately the same level of statistical precision

in the RRT condition as in the direct questioning condition (DQ), individuals were randomly assigned with a ratio of 2:1 (Warner 1965; Cohen 1988; Lensvelt-Mulders et al. 2005b). The unequal assignment to the experimental conditions is necessary due to the additional random noise component in the RRT.

Based on the administrative data, regression analyses were conducted for the gross sample, showing that randomization into experimental groups was successful. This approach resulted in 1,145 completes in the DQ condition and 2,066 in the RRT condition. Table 1 provides an overview of the assignments to the experimental conditions.

Of the respondents originally assigned to the RRT, 13.2 percent refused the application of the randomized response technique (DQ_RRT) and were subsequently asked to respond to the relevant survey questions directly ($n = 274$). Results from a multiple logistic regression model, not presented here, modeling refusal to comply with the randomization protocol (DQ_RRT) show that two variables in particular have a large, statistically significant effect and predictive power: poor language skills and whether a respondent refused to answer the question on household income both substantially increase the probability of a refusal. Refusal is also higher in the UB II sample, among younger and single respondents, among respondents who have never held a job before, and respondents with a lower socioeconomic status. Further analyses indicate that both splits do not differ with respect to gender, formal training, older age groups, a previous socially undesirable response, composition of social networks, various attitudes towards undeclared work – the focus of the original study – or region of residence. Given these results, all further analyses will also be conducted separately.

The survey instrument was fully standardized: All survey participants received identical instructions with respect to the voluntary nature of the survey, the survey topic, assurances of confidentiality and anonymity, definitions or further explanations regarding receipt or UB II if needed. The only differences are within the experimental splits.

Across the two samples, two different operationalizations were used: for the UB II sample – known to have received benefits in June 2010 – participants were asked to report any 'benefit receipt ever'. In the employee sample participants were asked to report receipt in 'September 2010'. While these different operationalizations guarantee that (aggregate) responses can be validated, another criterion was to keep the questions as simple as possible in order to ensure understanding and correct recall (Tourangeau et al. 2000; Groves et al. 2009; Manzoni et al. 2010). To ease recall in the employee sample (and allow validation), the question relates to a defined period of receipt just prior to data collection. Further, all question formats were kept as similar as possible to commonly used questions in labor market surveys (cf. the PASS study as described by Trappmann et al. 2010).

Table 1. Experimental conditions

Assigned condition	N	Realized condition	N
DQ	1,145	DQ	1,145
RRT_assigned	2,066	RRT	1,792
		DQ_RRT	274

2.1.3. The RRT Implementation

Lensvelt-Mulders et al. (2005b) compared the efficiency of various RRT designs. The authors demonstrate that one variant, the so-called forced-choice RRT variant (Boruch 1971), was shown to be among the statistically most efficient RRT designs, is usually well understood (Landsheer et al. 1999; De Schrijver 2012) and shows higher rates of rule compliance compared to other RRT designs (Böckenholt et al. 2009).

In the symmetric forced-choice design, respondents are instructed to reply according to a set of rules: the randomization device determines whether the respondent is forced to answer 'Yes' (with probability p_1) – irrespective of their true status –, 'No' (with probability p_2) – irrespective of their true status –, or whether the sensitive question is to be answered truthfully, that is 'Yes' or 'No' (with probability p_3). The survey was designed to minimize two respondent hazards: neither a positive nor a negative answer should risk suspicion. The advantage of this so-called 'symmetric' variant of the forced-choice RRT is that it is shown to reduce respondents incentive to cheat in the RRT condition (i.e., provide a negative response when they should say 'Yes') and leads to greater rule compliance compared to an asymmetric variant that protects only singular responses (Ostapczuk et al. 2009). Regarding statistical efficiency, Lensvelt-Mulders et al. (2005b) recommend that the probability of providing a forced 'Yes' should be approximately the same as the expected prevalence of the sensitive item under investigation (Clark and Desharnais 1998), while the probability to tell the truth should be between 0.7 and 0.8.

Assuming that the probability distribution of the randomization procedure is known, the population prevalence as well as standard errors (s.e.) and confidence intervals for the forced-choice RRT can be estimated as follows (Fox and Tracy 1986): the observed sampling distribution of 'Yes' responses $\hat{\Phi}$ is used as an estimator for the unknown population parameter Φ. The overall proportion of positive responses (Φ) is the sum of the proportion of 'forced Yes' responses (p_1), and the product of the (unknown) population parameter π multiplied by the probability of having to respond truthfully (p_3): $\Phi = p_1 + p_3 * \pi$. The prevalence estimate of the sensitive characteristic $\hat{\pi}_{RRT}$ is then given as (Lensvelt-Mulders et al. 2005b):

$$\hat{\pi}_{RRT} = \frac{\hat{\Phi} - p_1}{p_3} \tag{1}$$

An estimate of the sampling variance of $\hat{\pi}_{RRT}$ is given as:

$$Var(\hat{\pi}_{RRT}) = \frac{\hat{\Phi} * (1 - \hat{\Phi})}{n * (p_3)^2} \tag{2}$$

where n is the sample size.

Regarding the administration of the forced-choice RRT over the telephone, the RRT design as developed by Krumpal (2012) was implemented and refined based on results of pretest interviews (cognitive pretest $n = 31$; pretest with the fully programmed instrument $n = 63$). Krumpal (2012) demonstrates that those instructions are well understood by respondents and elicit more undesirable responses yielding higher prevalence estimates of xenophobia and anti-Semitism in Germany.

More precisely, in the survey respondents in the RRT condition were asked first to gather three coins, paper and pencil in order to note down the rules. Respondents were then asked to flip the three coins prior to each question in the RRT section. Should a respondent accidentally reveal the outcome of the coin flip, interviewers were trained to ask respondents to flip the coin again without revealing the outcome. The exact rules implemented to provide an answer were the following (for the entire instructions see Appendix, translated from German):

> *". . . please, answer as follows: 3 tails, please always respond with 'Yes;' 3 heads, please always respond with 'No;' a mixture, that is a combination of heads and tails, such as 2 heads and 1 tail, please respond truthfully."*

(Note to the reader: Interviewers were trained to leave enough time 1) for respondents to note down the rules and 2) for respondents to toss the coins and possibly to consult their notes.)

It follows from this that $p_1 = 0.125$ 'forced Yes,' $p_2 = 0.125$ 'forced No,' and $p_3 = 0.75$ truthful response. The main interest of the original study was 'undeclared work' (see Section 2), with an assumed prevalence of about 10 percent to 12 percent in Germany. The probabilities of a forced 'Yes'/'No' and 'the truth' were chosen accordingly. Regarding the above mentioned recommendations, this design is not optimal with respect to the investigation of UB II receipt.

To ensure respondent understanding of the technique as stressed in Landsheer et al. (1999), a minimum of one 'training' example – in which the true answer had been reported by the respondent earlier in the questionnaire – was provided to everyone in this experimental condition so as to familiarize the respondents with the RRT (for the implementation of the training example, see Appendix). If this 'training example' was answered incorrectly, or the interviewer was under the impression that the technique had not been fully understood, another standardized example was provided to the respondent. Only when full understanding of the rules had been assured, did the main RRT section begin.

2.1.4. Independent Variables and Operationalizations

A range of indicators explaining underreporting of UB II will be analyzed in the scope of the second research question. Existing empirical evidence shows that underreporting of UB II is more frequent among males, among people aged 25 and younger as well as employed people (Kreuter et al. 2014). The authors also find a significant effect of recall period and household size. Those respondents with a longer recall period and those living in a larger household underreported more frequently. Household size in this particular instance is not to be taken literally: rather it is an indicator capturing a higher propensity to conduct the interview with someone less knowledgeable about the receipt of UB II, and thus response error should be larger. Kreuter et al. (2010, 2014) also show that respondents who are more reluctant to participate in a survey are slightly more likely to underreport benefit receipt. The authors attribute this effect to a lower motivation of these respondents while controlling for sample composition and recall error due to a longer recall period. Both studies mentioned above only applied direct questioning techniques.

Drawing on main insights of these studies, as well as on behavioral theories and the response process (Tourangeau and Rasinski 1988), variables that capture subjective costs, risks and utilities that are associated with accurate reporting of UB II will be included in the models. It can be reasonably assumed that significant (negative) effects regarding reporting accuracy in the model of the direct questioning split are observed for those characteristics that are associated with higher subjective reporting costs. These are higher if receipt of UB II is perceived as particularly sensitive, for example, when a respondent is employed.

Table 2 presents an overview of all independent variables. Factors contributing to perceived item sensitivity and hence associated reporting costs, comprise: employment status, occupational status, and a respondent's willingness to provide socially undesirable answers. Further, the reluctance of the respondents to answer sensitive questions is operationalized with an indicator variable, measuring item nonresponse for the item household income. Equally important is a measure of how common the receipt of UB II is in a respondent's environment: admitting to receiving UB II could then be perceived as less of a norm violation and reported more accurately. Ideally this indicator would be measured at the neighborhood level, which is not possible in this particular case due to data privacy issues. Thus, the recipient rate at the more aggregate municipal level is included in all models.

According to the work of Böckenholt and van der Heijden (2007), the RRT works especially well if the RRT instructions are clearly understood and the cognitive burden is kept as low as possible. A second set of indicators thus relates to the survey process and to the application of the RRT by the respondents. The first indicator captures whether a respondent was reluctant to cooperate in the RRT condition (DQ_RRT) and was then surveyed in the direct questioning mode. In order to capture understanding of the RRT, two proxy indicators are used (Landsheer et al. 1999): first, interviewers were asked to rate the language skills (German) of a respondent immediately following the telephone interview. A second indicator pertaining to the understanding of the RRT instructions is educational attainment (formal training). Response latency, that is, the speed at which a respondent answers, is used as a measure for response quality.

All models control for gender (0 male, 1 female), age (below 25, 25–40, 41–57, 58 and above), which region of Germany a respondent resides in (0 West, 1 East) and single-person household (0 multi-member household, 1 single-person household). Including these controls seems appropriate given respondents refusing to stay in the assigned RRT condition and the assumed differential underlying mechanisms in both experimental groups.

2.2. *Register Data*

The analysis uses supplementary register data based on social security reports and reports from the FEA itself as gold standard. Information relating to basic income receipt is a by-product of the FEA activities, that is, process data generated from information provided by the applicants during the application process. This information, such as household composition or income, is used to evaluate entitlement to receive UB II. These de-identified basic income receipt records are accessible to researchers at IAB.

Table 2. Description of variables used in the multivariate analyses

Indicator	Description	
Factors contributing to perceived reporting costs and item sensitivity		
Employment status	At the time of survey	
	0	Not employed (unemployed, parental leave, student etc.)
	1	Marginally employed with income up to 400€
	2	Employed with labor income >400€
Occupational status	International socioeconomic index of occupational status (ISEI) (Ganzeboom et al. 1992). Coded based on ISCO88 of present or last job (Hendrickx 2002)	
	0	No ISEI available, that is, never held a job before (score =.)
	1	Low or medium ISEI of present or last job (score 16–43)
	2	High ISEI of present or last job (score >43)
Socially undesirable response	Socially undesirable response regarding tax honesty. Tax honesty is:	
	0	Absolutely worthwhile, worthwhile
	1	Not worthwhile, absolutely not worthwhile
Reluctance	Item nonresponse for household income	
	0	Substantive response
	1	Missing response
Recipient rate	Share of UB II in municipality	
Survey process and application of RRT		
RRT refusal (DQ_RRT)	0	RRT condition
	1	DQ_RRT condition
Language skills	Scale from 1= very good to 6= nonexistent (recoded 0,1)	
	0	Good (<3)
	1	Poor (>=3)
Formal training	0	Secondary degree and below
	1	Tertiary degree
Response latency	Standardized response time in experimental section (recoded according to quartiles)	
	0	Slow response ($<Q_{25}$)
	1	Mean response ($Q_{25}-Q_{75}$)
	2	Fast response ($>Q_{75}$)

For the analyses, only one indicator in these records is of relevance: whether an individual received UB II. As a general rule, all data relevant to payments and claims (taxes, pensions, unemployment benefits etc.), that is, the primary use of the social security system, are known to be of very good data quality (Jacobebbinghaus and Seth 2007). The analyses thus rest on the crucial assumption that the true value of the respondents can be

captured with these data. The UB II receipt indicator is known to be both accurate and complete and can serve as gold standard.

2.3. Combined Data

Since respondents were not asked for consent to link their survey data to the administrative data, the two data sources cannot be merged at an individual level.

However, due to sampling on the dependent variable (known as reverse record check studies; Groves 1989), each individual in the UB II sample should by default respond with 'Yes' to the 'benefit receipt ever' question. Overreporting is not possible by definition. With the true aggregate prevalence being 100 percent, an indicator variable can be created on the individual person level that captures whether an individual reported accurately without linkage of the two data sources. This measure of reporting accuracy is a binary variable that takes on the value 1 if the survey report matches the true value in the administrative records, and 0 if the survey report is 'No,' that is, a mismatch between the survey data and the administrative records. Item nonresponse is equally spread across all experimental conditions (three out of 1,598 respondents). Those cases are excluded from the analyses.

For the employee sample, the missing linkage consent question only allows an assessment of the first research question. Since it is not possible to link the survey data to the respective administrative records, it is impossible to construct a variable indicating reporting accuracy at an individual level. However, it is possible to derive and compare aggregate measures for respondents. According to the administrative data, the true aggregate prevalence of 'benefit receipt in September 2010' for respondents of the employee sample is 3.0 percent in the DQ condition and 4.2 percent for the RRT_assigned condition. Only the original assignment (DQ or RRT_assigned) and the response indicator can be used to obtain these true values. This is why RRT and DQ_RRT cannot be separated and have identical values. In the employee subsample, overreporting could theoretically be an issue. However, it seems unreasonable to assume that respondents, aside from overreporting due to satisficing or acquiescence (Krosnick 1991), would (consciously) overreport UB II receipt. Item nonresponse occurred once in the DQ condition (out of 1,613 respondents).

Due to the above mentioned limitations in the employee sample, the second research question can only be addressed using the UB II sample.

2.4. Statistical Analyses

The response bias is used to assess the impact of measurement error from the two alternative techniques of data collection. The bias of a statistic is simply the difference between the statistic's expectation and the true population value. The estimator of the response bias (B_j) in the respective experimental condition j is thus (adapted from Biemer 2010, 49):

$$B_j = \bar{y}_{j,svy} - \bar{y}_{j,adm} \tag{3}$$

which is the difference of the means of accurate reporting in the sample survey measurements ($\bar{y}_{j,svy}$) and the gold standard measurements ($\bar{y}_{j,adm}$). This approach will then

allow for a comparison of the overall response bias of the RRT and the DQ in both subsamples using one-sided unpaired t-test assuming unequal variances.

Subsequent to analyzing the overall bias for both samples (research question 1), logistic regression models will be used to model accurate reporting by experimental condition as a function of covariates for the UB II sample (research question 2). Again, the dependent variable Y_{ij} represents an individual's (i) response behavior (0 underreporting, 1 accurate reporting) in the experimental condition j. If the assumptions of privacy protection in the RRT condition hold, predictors related to perceived item sensitivity in the DQ model should be more positively related to accurate reporting. While for the direct condition a logistic regression model is appropriate, the RRT requires a logistic regression with an adapted likelihood function that accounts for the additional noise introduced by the RRT procedure, such as rrlogit (Jann 2011).

3. Empirical Results

Table 3 shows the prevalence estimates in percent for all experimental groups across both subsamples ($\bar{y}_{j,svy}$), the resulting response bias estimates (B_j in %pts) as well as the difference in biases ($B_{DQ} - B_j$ in %pts). Estimates presented in column 'RRT_assigned,' are based on the logic of 'intention-to-treat' analysis (Angrist et al. 1996). They provide a more conservative estimate of the average treatment effect of assignment. The last two columns take into account whether respondents actually received treatment – that is the RRT – or refused its application: 18 percent of the UB II respondents and 9 percent of the employee sample did not follow the randomization protocol. Since the exact questions asked in the survey differ across the two subsamples, response bias estimates are not comparable across subsamples and should be interpreted individually. The estimated response bias pointing in the expected direction is boldfaced, indicating a statistically significant amount of underreporting.

Replicating results from prior studies (Kreuter et al. 2010, 2014), receipt of benefit is underreported in both DQ conditions: for the UB II sample benefit receipt is underreported by 13.0 percentage points. While receipt of benefits is also underreported by 0.9 percentage points in the employed sample, this result is statistically nonsignificant. In absolute terms, the bias is larger in the UB II sample; in relative terms, standardized on the value of true prevalence, it is much larger in the employed sample (29.3% compared to 13.0%). However, these differences could be confounded by the fact that the question of receipt 'ever,' in the UB II sample, as opposed to 'September,' in the employee sample, might be perceived as less difficult or less sensitive by the respondents.

3.1. Reduction of Response Bias by Means of RRT?

Assuming that bias is solely due to item sensitivity and that the RRT can alleviate this bias, the RRT survey data estimates in Table 3 – granted that the RRT is understood and trusted – should not diverge significantly from the gold standard.

Contrary to the initial expectations, the response bias in the RRT_assigned condition differs significantly from zero. In the UB II sample, receipt of welfare benefits is underreported by 12.7 percentage points and, in the employee sample, by 1.9 percentage points. As for the DQ condition, the relative bias is larger in the employee sample (45.7%)

Table 3. (Estimated) proportions ($\bar{y}_j$), absolute response bias (B_j) and differential response bias ($B_{DQ} - B_j$)

Sample type	Estimate/statistic	DQ	RRT_assigned	RRT_assigned	
				RRT	DQ_RRT
UB II	$\bar{y}_{j,adm}$	1.000	1.000	1.000	1.000
	$\bar{y}_{j,svy}$	0.870	0.873	0.854	0.906
	B_j (s.e.)	**−0.130** (0.014)	**−0.127** (0.014)	**−0.146** (0.020)	**−0.094** (0.022)
	t-statistic	−9.274	−9.045	−7.465	−4.321
	$B_{DQ} - B_j$ (s.e.)		−0.003 (0.020)	0.016 (0.024)	−0.035 (0.026)
	t-statistic		−0.140	0.683	−1.353
	sample size (n)	579	1,016	836	180
	effective n	579	650	470	180
Employee	$\bar{y}_{j,adm}$	0.030	0.042	0.042	0.042
	$\bar{y}_{j,svy}$	0.021	0.023	0.004	0.043
	B_j (s.e.)	−0.009 (0.006)	−0.019 (0.014)	**−0.038** (0.014)	0.001 (0.021)
	t-statistic	−1.449	−1.393	−2.659	0.049
	$B_{DQ} - B_j$ (s.e.)		0.010 (0.015)	0.030 (0.016)	−0.010 (0.022)
	t-statistic		0.689	1.887	−0.447
	sample size (n)	564	1,048	955	93
	effective n	564	630	537	93

compared to the UB II sample (12.7%). Conducting separate analyses for those respondents who complied with the randomization protocol and those who did not, response bias for the RRT is larger for the former group in both samples (UB II sample: 14.6%pts vs. 9.4%pts; employee sample: 3.8%pts vs. 0.1%pts). Respondents who refused to apply the RRT are the ones who show the lowest levels of underreporting in both subsamples across all experimental conditions and thus seem to be the more accurate respondents (also in relative terms: 9.4% and 2.3%).

Furthermore, the RRT estimates should be less biased compared to those in the DQ condition ($B_{DQ} - B_j$ in %pts), resulting in a negative difference. The difference in response bias estimates in the UB II sample is statistically nonsignificant across all conditions: the response bias is 0.97 times smaller in the RRT_assigned condition compared to the DQ condition, 1.13 times higher for RRT and 0.73 times smaller for DQ_RRT. In the employee sample, the differences are nonsignificant as well: the response bias is 2.12 times higher in the RRT_assigned condition compared to direct questioning and 0.12 times smaller for DQ_RRT. Contrary to the expectations, it is significantly larger in the RRT condition (4.35, $p = 0.03$).

To summarize some of the results for the initial research question: 1) the particular forced-choice telephone implementation of the RRT cannot reduce bias in the estimated prevalence of basic income support in Germany, while 2) the RRT performs significantly worse if the item under investigation is of a low prevalence rate, as in the case of the employee sample. Furthermore, due to the random noise in the RRT condition, variance estimates are inflated by a factor of $1.\bar{7}$ or, put differently, the effective sample size is reduced accordingly. All other things being equal, this leads to an increased mean squared error (MSE) in the RRT condition. The MSE estimate in the UB II DQ condition is 0.13 and 0.02 in the employee sample. Assuming identical sample sizes in both conditions, namely those of the respective DQ split, the MSE in the UB II RRT condition would then be 0.28 and 0.19 in the employee sample. Since the actual sample size in the RRT splits is larger, MSE estimates are 0.20 in the UB II sample and 0.11 in the employee sample.

One can only speculate about the reasons for the poor performance of the RRT in this particular study. One reason might be that the initial assumption – that unemployment benefit receipt is sensitive – is false. In that case, one would not expect to see the RRT producing estimates closer to the truth compared to direct questioning. The second argument might be that respondents do not apply the randomization procedure correctly, that is, that either they do not flip coins at all or they do not adhere to the RRT instructions (Clark and Desharnais 1998). In the first instance this could mean that a face-to-face implementation, with an interviewer supervising the randomization procedure, could perform better. The second issue is trust in the method: despite understanding the method, it is also crucial that respondents trust the privacy protection provided by the RRT (Holbrook and Krosnick 2010; Coutts and Jann 2011). While it can be reasonably assumed that unintentional noncompliance with the rules, that is, respondents accidentally providing a wrong answer, should not occur if the method is understood, nevertheless trust is essential. Respondents might consciously decide to edit their answers and ignore the RRT instructions if they lack trust: they might respond 'No' even if the randomization device prompted them to answer 'Yes' (cheating). Or, if prompted to answer truthfully, respondents might edit their answer and report a 'No' (even if the truth is 'Yes'), resulting

in underreporting. These so-called 'cheaters' and 'under reporters' lead to the fact that the RRT estimates are biased (see also Boeije and Lensvelt-Mulders 2002; Böckenholt et al. 2009; Coutts and Jann 2011; Ostapczuk et al. 2011). Specific to our study, there is a unique, indirect method of assessing the amount of cheating and underreporting relying on a few assumptions: 1) overreporting (incl. false positives in the employee sample) does not occur, 2) both effects are homogeneous across samples. Equation $\Phi_{sample} = p_1*(1-c) + p_3*(1-u)*\pi$ introducing a cheating as well as an underreporting parameter is then identifiable. Estimates of cheating in the RRT condition of this study amount to 18.4 percent and underreporting of 11.5 percent. These results also underline the utility and necessity of designs that allow for an estimation of and correction for cheating (Clark and Desharnais 1998; Böckenholt et al. 2009; Van den Hout et al. 2010; Ostapczuk et al. 2011; De Jong et al. 2012). These designs typically allow the identification and estimation of a cheating parameter by assigning two different misclassification probabilities to different RRT subsamples. The particular design of this study was chosen due to a successful prior implementation in the study conducted by Krumpal (2012), considerations of a loss in statistical efficiency and the proposed indirect estimation strategy. A third reason for the poor performance of the RRT could be the mode of data collection via telephone itself. Respondents might find it easier to 'cheat' on the phone than in a face-to-face mode (De Leeuw and van der Zouwen 1988; Aquilino 1994).

This result is particularly relevant for future studies due to the cost implications: the increased costs in the RRT condition are due to – all other things being equal – a larger sample size, longer interview durations (the RRT section was on average six minutes longer than DQ; see also Wolter and Preisendörfer 2013), statistically more complex analyses, more intensive interviewer training and, most important, a higher respondent burden. Given the empirical evidence, the additional costs of a forced-choice RRT data collection for welfare receipt are not justified. Thus, in terms of bias versus efficiency, these results clearly favor direct questioning to collect data on welfare benefit receipt in Germany.

3.2. *Is Response Bias Subgroup Specific?*

Contrary to the expectations in both experimental conditions, the results for research question 1 indicate a tremendous amount of misreporting.

The following section will analyze response error between subgroups while controlling for a differential sample composition across both experimental conditions. Since individual-level data is available only for the UB II sample, further analyses are limited to this sample and inferences can only be drawn with respect to this specific population. The dependent variable, 'accurate reporting,' will be modeled separately as a function of several individual characteristics for respondents in the UB II sample for each experimental split. In order to account for potential nonlinear relationships, all variables enter the regression equation categorically.

Table 4 displays the average marginal effects (AME) from logistic regression models (Stata version 12.1, rrlogit, Jann 2011), modeling accurate reporting as a function of the covariates mentioned above, as well as the difference in AMEs (*DQ − RRT_assigned*). The AME is the average of discrete or partial changes over all observations. It yields a

Table 4. Logistic regression models analyzing accurate reporting of receipt of UB II (average marginal effects and 95% confidence intervals)

Y: Accurate reporting	Model 1: DQ	Model 2: RRT_as.	Difference: DQ − RRT_as.
Factors contributing to perceived reporting costs and item sensitivity			
No employment (ref. employed >400€)	**0.118*** **	**0.095** **	0.023
	[0.055;0.180]	[0.025;0.165]	[−0.071;0.117]
Marginally employed	0.017	0.013	0.004
	[−0.050;0.083]	[−0.056;0.082]	[−0.092;0.100]
Low/Med. ISEI (ref. n/a (never employed))	**0.069**$^{+}$	**0.063**$^{+}$	0.006
	[−0.000;0.137]	[−0.010;0.136]	[−0.094;0.106]
High ISEI	0.037	−0.003	0.040
	[−0.046;0.120]	[−0.098;0.092]	[−0.086;0.167]
Socially undesirable response (tax honesty)	**−0.045**$^{+}$	**0.158** **	**−0.203** **
	[−0.097;0.007]	[0.039;0.276]	[−0.333;−0.074]
Reluctance (item NR)	**−0.113** **	**−0.148*** **	0.036
	[−0.186;−0.039]	[−0.227;−0.069]	[−0.073;0.144]
Recipient rate	0.050	0.069	−0.018
	[−0.221;0.322]	[−0.241;0.379]	[−0.043;0.394]
Survey process and application of RRT			
DQ_RRT	–	0.049	–
	–	[−0.020;0.117]	–
Language skills (poor)	**−0.057**$^{+}$	−0.058	0.000
	[−0.120;0.005]	[−0.128;0.013]	[−0.094;0.095]
Tertiary degree	0.028	0.075	0.047
	[−0.072;0.128]	[−0.042;0.192]	[−0.201;0.107]
Fast response (ref. mean response)	0.038	−0.004	0.042

Table 4. Continued

Y: Accurate reporting	Model 1: DQ	Model 2: RRT_as.	Difference: DQ – RRT_as.
	[−0.030;0.106]	[−0.070;0.061]	[−0.052;0.137]
Slow response	0.011	0.042	−0.031
	[−0.049;0.071]	[−0.030;0.114]	[−0.125;0.062]
Controls			
Female	−0.001	**0.049**$^{+}$	**−0.050**$^{+}$
	[−0.051;0.049]	[−0.005;0.103]	[−0.124;0.023]
Age <25 (ref. 25 to 40)	**−0.137*****	**−0.168*****	0.031
	[−0.199;−0.076]	[−0.240;−0.097]	[−0.063;0.125]
Age 41 to 57	0.007	−0.028	0.035
	[−0.055;0.069]	[−0.109;0.053]	[−0.067;0.138]
Age >57	−0.008	−0.063	0.054
	[−0.122;0.105]	[−0.164;0.039]	[−0.098;0.207]
East Germany	0.009	0.032	−0.023
	[−0.050;0.069]	[−0.037;0.102]	[−0.114;0.068]
Single-person household	**0.092***	**0.071***	0.021
	[0.020;0.163]	[0.001;0.140]	[−0.079;0.121]
Model fit			
N	579	1,016	
LR Chi2 (df)	119.98 (17)	95.89 (18)	
Pseudo R2	0.27	0.09	
AIC	360.43	966.08	
BIC	434.58	1054.70	

95% confidence intervals in brackets; $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

straightforward interpretation of estimation results and effect sizes, and allows a comparison between models (Bartus 2005; Mood 2010). Subsequently, only AMEs will be reported.

Two models are presented in Table 4: Model 1 analyzes accurate reporting in the direct questioning condition and serves as a baseline for examining reporting accuracy. Model 2 replicates the same model in the RRT condition. I expect to see more accurate reporting in the RRT condition, especially for those variables related to perceived item sensitivity. Thus all (negative) effects related to item sensitivity that are found in the direct split should become more positively (or nonsignificantly) related to accurate reporting. This second model also presents insights regarding the question of which variables related to the survey process contribute to more accurate reporting.

Turning to the DQ Model 1, those variables related to perceived item sensitivity are of particular interest. Unconditional on other covariates, as expected, respondents with no current *employment* are on average 11.8 percentage points more likely than respondents with an income of 400 Euro and above to report receipt of UB II. Marginally employed respondents do not differ systematically from the reference category. Regarding *occupational status* respondents with a high (present or past) status are expected to report receipt of UB II less often than the other categories. Contrary to the initial expectations, respondents with a high ISEI have a slight tendency to report more accurately compared to the reference category (no job), while respondents with a low or medium status report receipt significantly more accurately (6.9%pts) than those who have never held a job before. Regarding the difference between respondents with a high ISEI and those with a low or medium ISEI, no significant difference is observed. The item '*socially undesirable response*' regarding tax honesty significantly explains accurate reporting, but in a surprising way: respondents with an honest, but more socially undesirable attitude towards tax dishonesty are on average 4.5 percentage points more likely to underreport the receipt of basic income support than those respondents displaying a more desirable attitude towards tax honesty. At first, this finding seems counterintuitive: responding in a socially undesirable manner in one instance would result in a higher propensity to admit another undesirable characteristic. One potential explanation could be that, given that 'tax dishonesty' is acceptable, misreporting on other characteristics is considered acceptable as well. *Reluctance* contributes significantly to the explanation of underreporting of UB II (11.3%pts). Regarding the *share of UB II* recipients at the municipality level, there is no significant effect, supporting the hypothesis regarding the wrong level of measurement.

Those characteristics relating to the survey process contribute less to the explanation of accurate reporting. Poor *language skills* are the only significant predictor contributing to underreporting of UB II (5.7%pts). With respect to the controls, younger respondents, aged 24 and below, significantly underreport receipt (13.7%pts). In line with expectations, the indicator 'single-person household' significantly improves reporting accuracy (9.2%pts). Both results support the argument that proxy reports with less knowledgeable persons on receipt of UB II are less accurate, since younger respondents are more likely to still live with their parents who apply for UB II for the entire household.

Turning to Model 2—the RRT model—the results are strikingly similar, both in direction and magnitude. Contrary to the expectations, variables related to perceived item

sensitivity exert approximately the same influence as in the DQ model with one exception: *socially undesirable response*. Respondents stating that tax honesty is (absolutely) not worthwhile report on average 15.8 percent more accurately in the RRT condition. This difference between both models is statistically significant, indicating that the RRT reduces social desirability concerns for those respondents ($p < 0.01$). Given this evidence, the above explanation for this finding seems implausible. A different explanation might help to solve the puzzle: in Germany, tax dishonesty is largely associated with undeclared work/income. Receipt of UB II is based on accurate reporting of all forms of income and misreporting of income to the authorities is heavily pursued. Stating that tax honesty is (absolutely) not worthwhile in the direct questioning condition might be considered indirect evidence for potential concealing of income when applying for UB II and is thus a highly sensitive question itself when confirming receipt of UB II. This would explain the negative relationship. This same question is potentially perceived as less intrusive in the RRT condition and hence respondents more openly state their opinion. The positive relationship in Model 2 is thus internally consistent.

To summarize, contrary to expectation, the RRT does not elicit more accurate reports for respondents for whom reports of UB II can be assumed to be particularly sensitive, with one exception. This indicates that the same misreporting mechanisms are at work in both experimental conditions.

Similar to Model 1, those characteristics relating to the survey process and the application of the RRT overall contribute less to the explanation of accurate reporting. Respondents who *refused the application of the RRT* report more accurately than those respondents in the RRT condition (4.9%pts). Anecdotal evidence from interviewer observations suggests that those respondents either distrust the RRT or claim that they 'have nothing to hide' and want to be questioned directly. The effect size of lack of *language skills* is negative and roughly the same as in Model 1; however, it just fails to be statistically significant ($p = 0.101$). It can be assumed that respondents who do not accurately understand what is asked of them in either condition (particularly so in the RRT) will not trust the method and therefore report (a 'self-protective' or 'nonincriminating') 'No' (Böckenholt et al. 2009; Coutts and Jann 2011). Thus the result is as expected for both models. Remember that while a *tertiary degree* contributes to accurate reporting (2.8%pts) in Model 1, in Model 2 this effect is larger in comparison to Model 1, but not compared to the reference category (7.5%pts). Due to the small number of people holding a tertiary degree, confidence intervals are rather large for this estimate. Further regression analyses were conducted but are not presented here: they account for the fact that if language skills are poor, neither educational degree will make a difference in the reporting accuracy. Assuming good language skills (essentially modeling an interaction), the results show a larger effect of university degree in Model 2. This suggests that the RRT reduces underreporting for these respondents: however, it remains unclear whether this effect is due to a better understanding of the RRT compared to the reference category (Poor German Skills and No Tertiary Degree) or the RRT guaranteeing anonymity and reducing item sensitivity for the more highly-educated group. *Response latency*, that is, the speed at which a respondent answers, is used as a measure for response quality. Surveying in the RRT condition by definition takes longer than a comparable direct question, since

respondents have to follow the RRT protocol. In theory, irrespective of the experimental condition, a longer answering process could indicate more editing of the true response and thus a poorer data quality (Holtgraves 2004). On the other hand, it could also be associated with higher-quality information and processing in the RRT condition (Wolter 2012). Results for response latency exhibit no clear pattern across models and are nonsignificant: in Model 2, a slower response indicates on average greater accuracy (4.2%pts; 0.4%pts more underreporting for fast respondents; this difference is statistically nonsignificant), while in Model 1, both fast and slow reporting is associated with greater accuracy compared to the reference category (3.8%pts and 1.1%pts).

With respect to the controls, effects are similar to those of Model 1, with the exception of women on average reporting more accurately in Model 2 (4.9%pts). The difference between both models is statistically significant ($p < 0.10$).

To summarize the results, results from previous studies (Kreuter et al. 2014) can be replicated in Model 1, that is, especially for characteristics relating to item sensitivity (employment status, occupational status, socially undesirable response, reluctance) and structural characteristics (age, single-person household). Contrary to the initial expectations, the RRT cannot resolve social desirability concerns for these items; as expected, structural influences persist. The hypotheses relating to the survey process and the application of the RRT cannot be confirmed with these results.

Analyzing DQ, RRT and DQ_RRT in one joint model while controlling for covariates shows that while RRT and DQ_RRT result in more accurate responses, these effects are statistically nonsignificant. A fully interacted model (all covariates and the RRT indicator) yields the following significant interaction effects: more accurate reporting by respondents with a socially undesirable response and those with a tertiary degree, as well as respondents taking longer to respond under RRT.

4. Discussion and Conclusion

The initial research question addressed the performance of a forced-choice telephone implementation of the RRT for the estimation of welfare receipt compared to direct questioning. The results show that this particular RRT design does not reduce underreporting in the data collection on welfare benefit receipt in a telephone survey. The RRT performs worse in the employee sample, where the overall prevalence is close to zero.

Insights into who underreports receipt of UB II were the main focus of the second research question. Inferences are limited to the population of UB II recipients in Germany. Reporting accuracy is significantly higher in both methods for respondents who perceive reporting of UB II as less of a norm violation, that is, respondents who are not employed. Respondents who admit to tax dishonesty report more accurately in the RRT model, but less accurately in the DQ model, as do respondents who are unwilling to provide information on other items such as income. Thus, there is a tendency for underreporting whenever receipt of welfare benefits is perceived as more sensitive in both models. If the RRT were to resolve the concerns of social desirability, differential effects would have been observed across both methods for those items capturing sensitivity. The results do not

support this argument: differences between models are statistically significant only for those respondents having given another socially undesirable response. Furthermore, it was expected that those items fostering understanding of the RRT would contribute to a higher reporting accuracy. While most effects point in the expected direction, they are statistically nonsignificant.

One can only speculate about the potential reasons for the failure of the RRT in this study. One argument discussed above relates to the potential lack of sensitivity of the item under study. If underreporting were not caused by perceived sensitivity, then the RRT would not be expected to decrease bias. Studies regarding the perception of welfare receipt would not support this argument (Bullock 2006). Other arguments explaining the poor performance of the RRT relate to 'cheating' and 'noncompliance' with the instructions of the RRT (Clark and Desharnais 1998; Böckenholt et al. 2009; De Jong et al. 2012). For one, it remains unclear whether respondents are really implementing the randomization procedure while on the telephone (Holbrook and Krosnick 2010). In that instance, a face-to-face mode might seem more appropriate. A second concern – which is more in line with the results – is that respondents 'forced' by the randomization device to provide a (false) positive answer might decide not to comply with the RRT rules (and reply 'No' instead of 'Yes') or underreport if asked to provide a truthful response (Böckenholt and van der Heijden 2007; Coutts and Jann 2011). This concern cannot be ruled out even in the face-to-face mode. However, it highlights the importance of RRT designs allowing for an estimation of underreporting and cheating, as prevalence estimates can then be corrected. The last argument pertains to the telephone mode itself: if the benefits of noncompliance are large and social control is weak, persons are less willing to comply (Böckenholt and van der Heijden 2007).

Overall, the finding that the (forced-choice variant of the) RRT still contains response bias has been confirmed by other recent studies (Holbrook and Krosnick 2010; Coutts and Jann 2011; Wolter and Preisendörfer 2013; Höglinger et al. 2014), but it is worth reiterating that the RRT does not outperform direct questioning (Lensvelt-Mulders et al. 2005). On the contrary, yielding approximately the same bias, mean squared error increased due to an inflated variance. Furthermore, there is a tremendous amount of visible refusal to follow the randomization protocol in the RRT condition as well as a large share of covert misreporting. Using this implementation of the RRT, the main implication is that the additional burden imposed on respondents in combination with additional surveying costs, for example in terms of sample size and duration, are not justified. Given that respondent burden is associated with a decreased probability of future survey participation and an increase in breakoffs, these results are particularly important. Overall, 95 out of 229 respondents broke off the interview during the RRT introduction or first item within the experimental section, while most of the 46 breakoffs in the DQ condition occurred either before or after the experimental condition, and none while asking about welfare benefits or undeclared work.

The evidence in this study also supports the notion that this particular RRT design performs slightly better in certain populations: those respondents with good language skills, those more highly educated, and those who take enough time to respond in the RRT condition, that is, the correct application of the randomization process being observed in some way. Furthermore, language skills and respondent reluctance are significant

predictors of whether respondents comply with the randomization protocol. When the research focus is on populations with a lower educational background, the results may thus be very different. The results and the tremendous amount of underreporting do not support the use of this implementation of the RRT in large-scale population surveys. Other techniques, such as the crosswise or triangular technique, a different variant of the RRT (Yu et al. 2008), might be a preferable method. These methods do not require a randomization device, are less of a cognitive burden for respondents, are easier to implement over the telephone, provide less incentives to misreport and might thus be a viable alternative to direct questioning (Jann et al. 2012; Korndörfer et al. 2014; Höglinger et al. 2014).

Appendix: RRT Introduction and Training Example

"I will now introduce you to a technique, that will allow you to keep your personal experiences anonymous by means of a coin flip. Even if this might sound strange to you, I kindly ask you to help us to try this new method. This method is scientifically approved and is fun. Would you please get a paper, a pencil, and three coins?

You will be able to answer all of the following questions either with 'Yes' or 'No.' Before answering each question, I would kindly ask you to flip the three coins. Please do not tell me the outcome of this coin flip. According to the outcome, please answer as follows:

- 3 tails; please always respond with 'Yes'
- 3 heads; please always respond with 'No'
- a mixture; that is, a combination of heads and tails, such as 2 heads and 1 tail, please respond truthfully

As you can see chance decides whether you actually respond to the question or provide a surrogate answer. Thus, your privacy is always protected. I, as the interviewer, will never know the result of your coin toss. Thus, I can never know, why you respond with 'Yes' or 'No.' Do you have any further questions regarding the technique?

Let us walk through one example together.

If you flip 3x heads, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. 'No,' according to the rule)*

If you flip 3x tails, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. 'Yes,' according to the rule)*

If you have a mixed result, for example, flip 2x heads and 1x tail, and I ask you if you are 18 years or older, what would you reply? *(Int: Pause; let the respondent reply first. The response has to be 'Yes' as part of the requirements of the sampling design)*

Do you have any further questions?"

(Note to the reader: If there were further questions, the rules were repeated and a new example provided before asking one question on UB II receipt followed by two questions on undeclared work.) (Translated from German)

5. References

AAPOR - The American Association for Public Opinion Research 2011. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th Ed. Lanexo: AAPOR.

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

Aquilino, W.S. 1994. "Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment." *Public Opinion Quarterly* 58: 210–240. DOI: http://dx.doi.org/10.1086/269419.

Bartus, T. 2005. "Estimation of Marginal Effects Using Margeff." *The Stata Journal* 5: 309–329.

Biemer, P.P. 2010. "Overview of Design Issues: Total Survey Error." In *Handbook of Survey Research*, edited by P.P. Biemer, P.V. Marsden, and J.D. Wright, 27–57. Bingley: Emerald Publishing Group Limited.

Boeije, H. and G.J.L.M. Lensvelt-Mulders. 2002. "Honest by Chance: A Qualitative Interview Study to Clarify Respondents' (Non-)compliance with Computer-Assisted-Randomized Response." *Bulletin de Methodologie Sociologique* 75: 24–39.

Boruch, R.F. 1971. "Assuring Confidentiality of Responses in Social Research: A Note on Strategies." *The American Sociologist* 6: 308–311.

Bradburn, N., S. Sudman, and B. Wansink. 2004. *Asking Questions. Revised Edition*. San Francisco: Jossey-Bass.

Bullock, H.E. 2006. "Attributions for Poverty: A Comparison of Middle-Class and Welfare Recipient Attitudes." *Journal of Applied Social Psychology* 29: 2059–2082. DOI: http://dx.doi.org/10.1111/j.1559-1816.1999.tb02295.x.

Böckenholt, U., S. Barlas, and P.G.M. van der Heijden. 2009. "Do Randomized-Response Designs Eliminate Response Biases? An Empirical Study of Non-Compliance Behavior." *Journal of Applied Econometrics* 24: 377–392. DOI: http://dx.doi.org/10.1002/jae.1052.

Böckenholt, U. and P.G.M. van der Heijden. 2007. "Item Randomized-Response Models for Measuring Noncompliance: Risk-Return Perceptions, Social Influences, and Self-Protective Responses." *Psychometrika* 72: 245–262. DOI: http://dx.doi.org/10.1007/s11336-005-1495-y.

Cialdini, R.B. 2007. "Descriptive Social Norms as Underappreciated Sources of Social Control." *Psychometrika* 72: 263–268. DOI: http://dx.doi.org/10.1007/s11336-006-1560-6.

Clark, S.J. and R.A. Desharnais. 1998. "Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model." *Psychological Methods* 3: 160–168. DOI: http://dx.doi.org/10.1037/1082-989X.3.2.160.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Vol. 2. Hillshale, NJ: Erlbaum.

Coutts, E. and B. Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count

Technique (UCT)." *Sociological Methods & Research* 40: 169–193. DOI: http://dx.doi.org/10.1177/0049124110390768.

Coutts, E., B. Jann, I. Krumpal, and A.-F. Näher. 2011. "Plagiarism in Student Papers: Prevalence Estimates Using Special Techniques for Sensitive Questions." *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)* 231: 749–760.

De Jong, M.G., R. Pieters, and S. Stremersch. 2012. "Analysis of Sensitive Questions Across Cultures: An Application of Multigroup Item Randomized Response Theory to Sexual Attitudes and Behavior." *Journal of Personality and Social Psychology* 19: 153–176. DOI: http://dx.doi.org/10.1037/a0029394.

De Leeuw, E.D. and J. van der Zouwen. 1988. "Data Quality in Telephone and Face to Face Surveys: A Comparative Metaanalysis." In *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg, 283–299. New York: John Wiley & Sons, Inc.

De Schrijver, A. 2012. "Sample Survey on Sensitive Topics: Investigating Respondents' Understanding and Trust in Alternative Versions of the Randomized Response Technique." *Journal of Research Practice* 8: 1–17.

Fidler, D.S. and R.E. Kleinknecht. 1977. "Randomized Response versus Direct Questioning: Two Data-Collection Methods for Sensitive Information." *Psychological Bulletin* 84: 1045–1049. DOI: http://dx.doi.org/10.1037/0033-2909.84.5.1045.

Fox, J.A. and P.E. Tracy. 1986. *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills: Sage Publications.

Ganzeboom, H.B.G., P.M. De Graaf, and D.J. Treiman. 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21: 1–56. DOI: http://dx.doi.org/10.1016/0049-089X(92)90017-B.

Greenberg, B.G., A.L.A. Abul-Ela, W.R. Simmons, and D.G. Horvitz. 1969. "The Unrelated Question Randomized Response Model: Theoretical Framework." *Journal of the American Statistical Association* 64: 520–539. DOI: http://dx.doi.org/10.1080/01621459.1969.10500991.

Greenberg, B.G., R.R. Kuebler Jr., J.R. Abernathy, and D.G.G. Horvitz. 1971. "Application of the Randomized Response Technique in Obtaining Quantitative Data." *Journal of the American Statistical Association* 66: 243–250. DOI: http://dx.doi.org/10.1080/01621459.1971.10482248.

Groves, R.M. 2004 [1989]. *Survey Error and Survey Costs*. Hoboken: Wiley & Sons.

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

Hausman, J. 2001. "Mismeasured Variables in Econometric Analysis: Problems From the Right and Problems from the Left." *The Journal of Economic Perspectives* 15: 57–67.

Hendrickx, J. 2002. "ISKO: Stata Module to Recode 4 Digit ISCO-88 Occupational Codes, Statistical Software Components s425802." Boston College Department of Economics. revised 20 Oct 2004. Available at: https://ideas.repec.org/c/boc/bocode/s425802.html (accessed February 14, 2015).

Holbrook, A.L., M.C. Green, and J.A. Krosnick. 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://dx.doi.org/10.1086/346010.

Holbrook, A.L. and J.A. Krosnick. 2010. "Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling into Question the Method's Validity." *Public Opinion Quarterly* 74: 328–343. DOI: http://dx.doi.org/10.1093/poq/nfq012.

Holtgraves, T. 2004. "Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding." *Personality and Social Psychology Bulletin* 30: 161–172. DOI: http://dx.doi.org/10.1177/0146167203259930.

Horvitz, D.G., B.V. Shah, and W.R. Simmons. 1967. "The Unrelated Question Randomized Response Model." In Proceedings of the Social Statistics Section. American Statistical Association, 65–72.

Höglinger, M., B. Jann, and A. Diekmann. 2014. *Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model*. University of Bern Social Science Working Paper No. 9, 1–51. Available at: ftp://repec.sowi.unibe.ch/files/wp9/hoeglinger-jann-diekmann-2014.pdf (accessed September 17, 2014).

Jacobebbinghaus, P. and S. Seth. 2007. "The German Integrated Employment Biographies Sample IEBS." *Schmollers Jahrbuch* 127: 335–342.

Jann, B. 2011. "Rrlogit: Stata module to estimate logistic regression for randomized response data." Statistical Software Components, Boston College Department of Economics. Available at: https://ideas.repec.org/c/boc/bocode/s456203.html (accessed February 14, 2015).

Jann, B., J. Jerke and I. Krumpal. 2012. "Asking Sensitive Questions Using the Crosswise Model. An Experimental Survey Measuring Plagiarism." *Public Opinion Quarterly* 71: 32–49. DOI: http://dx.doi.org/10.1093/poq/nfr036.

Kirchner, A. 2014. Techniques for Asking Sensitive Question in Labor Market Surveys. IAB-Bibliothek Dissertationen, 348. Bielefeld: Bertelsmann. Available at: http://edoc.ub.uni-muenchen.de/17192/1/Kirchner_Antje.pdf (accessed February 14, 2015).

Kirchner, A., I. Krumpal, M. Trappmann, and H. von Hermanni. 2013. "Messung und Erklärung von Schwarzarbeit in Deutschland – Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit." *Zeitschrift für Soziologie* 42: 291–314.

Korndörfer, M., I. Krumpal, and S.C. Schmukle. 2014. "Measuring and Explaining Tax Evasion: Improving Self-Reports Using the Crosswise Model." *Journal of Economic Psychology* 45: 18–32. DOI: http://dx.doi.org/10.1016/j.joep.2014.08.001.

Kreuter, F., G. Müller, and M. Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74: 880–906. DOI: http://dx.doi.org/10.1093/poq/nfq060.

Kreuter, F., G. Müller, and M. Trappmann. 2014. "A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents." Sociological Methods and Research 43: 452-464. DOI: http://dx.doi.org/10.1177/0049124113508094.

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

Krumpal, I. 2012. "Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning." *Social Science Research* 41: 1387–1403. DOI: http://dx.doi.org/10.1016/j.ssresearch.2012.05.015.

Kuk, A.Y.C. 1990. "Asking Sensitive Questions Indirectly." *Biometrika* 77: 436–438. DOI: http://dx.doi.org/10.1093/biomet/77.2.436.

Lamb, C.W. and D.E. Stem. 1978. "An Empirical Validation of the Randomized Response Technique." *Journal of Marketing Research* 15: 616–621.

Landsheer, J.A., P.G.M. van der Heijden, and G. van Gils. 1999. "Trust and Understanding. Two Psychological Aspects of Randomized Response. A Study of a Method for Improving the Estimate of Social Security Fraud." *Quality & Quantity* 33: 1–12. DOI: http://dx.doi.org/10.1023/A:1004361819974.

Lara, D., S.G. García, C. Ellertson, C. Camlin, and J. Suárez. 2006. "The Measure of Induced Abortion Levels in Mexico Using Random Response Technique." *Sociological Methods & Research* 35: 279–301. DOI: http://dx.doi.org/10.1177/0049124106290442.

Lara, D., J. Strickler, C.D. Olavarrieta, and C. Ellertson. 2004. "Measuring Induced Abortion in Mexico." *Sociological Methods & Research* 32: 529–558. DOI: http://dx.doi.org/10.1177/0049124103262685.

Lee, R.M. 1993. *Doing Research on Sensitive Topics*. London: Sage.

Lensvelt-Mulders, G.J.L.M., J.J. Hox, P.G.M. van der Heijden, and C.J.M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods & Research* 33: 319–348. DOI: http://dx.doi.org/10.1177/0049124104268664.

Lensvelt-Mulders, G.J.L.M., J.J. Hox, and P.G.M. Van der Hejden. 2005b. "How to Improve the Efficiency of Randomized Response Designs." *Quality & Quantity* 39: 253–265. DOI: http://dx.doi.org/10.1007/s11135-004-0432-3.

Lensvelt-Mulders, G.J.L.M., P.G.M. Van der Heijden, O. Laudy, and G. van Gils. 2006. "A Validation of Computer-Assisted Randomized Response Survey to Estimate the Prevalence of Undeclared Work in Social Security." *Journal of the Royal Statistical Society (Series A)* 169: 305–318.

Locander, W., S. Sudman, and N. Bradburn. 1976. "An Investigation of Interview Method. Threat and Response Distortion." *Journal of the American Statistical Association* 71: 269–275. DOI: http://dx.doi.org/10.1080/01621459.1976.10480332.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Mangat, N.S. 1994. "An Improved Randomized Response Strategy." *Journal of the Royal Statistical Society (Series B)* 56: 93–95.

Mangat, N.S. and R. Singh. 1990. "An Alternative Randomized Response Procedure." *Biometrika* 77: 439–442. DOI: http://dx.doi.org/10.1093/biomet/77.2.439.

Manzoni, A., J.K. Vermunt, R. Luijkx, and R. Muffels. 2010. "Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error." *Sociological Methodology* 40: 39–73. DOI: http://dx.doi.org/10.1111/j.1467-9531.2010.01230.x.

Mood, C. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About it." *European Sociological Review* 26: 67–82. DOI: http://dx.doi.org/10.1093/esr/jcp006.

Moors, J.J.A. 1971. "Optimization of the Unrelated Randomized Response Model." *Journal of the American Statistical Association* 66: 627–629. DOI: http://dx.doi.org/10.1080/01621459.1971.10482320.

Moshagen, M., E.B. Hilbig, E. Erdfelder, and A. Moritz. 2014. "An Experimental Validation Method for Questioning Techniques That Assess Sensitive Issues." *Experimental Psychology* 61: 48–54. DOI: http://dx.doi.org/10.1027/1618-3169/a000226.

Ostapczuk, M., M. Moshagen, Z. Zhao, and J. Musch. 2009. "Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry." *Journal of Educational and Behavioral Statistics* 43: 267–287. DOI: http://dx.doi.org/10.3102/1076998609332747.

Ostapczuk, M., J. Musch, and M. Moshagen. 2011. "Improving Self-Report Measures of Medication Non-Adherence Using a Cheating Detection Extension of the Randomized-Response Technique." *Statistical Methods in Medical Research* 20: 489–503. DOI: http://dx.doi.org/10.1177/0962280210372843.

Tourangeau, R. and K.A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103: 299–314. DOI: http://dx.doi.org/10.1037/0033-2909.103.3.299.

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

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133: 859–883. DOI: http://dx.doi.org/10.1037/0033-2909.133.5.859.

Tracy, P.E. and J.A. Fox. 1981. "The Validity of Randomized Response for Sensitive Measurements." *American Sociological Review* 46: 187–200.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt. 2010. "PASS: A Household Panel Survey for Research on Unemployment and Poverty." *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften* 130: 609–622.

Umesh, U.N. and R.A. Peterson. 1991. "A Critical Evaluation of the Randomized Response Method: Applications, Validation, and Research Agenda." *Sociological Methods & Research* 20: 104–138. DOI: http://dx.doi.org/10.1177/0049124191020001004.

Van den Hout, A., U. Böckenholt, and P.G.M. van der Heijden. 2010. "Estimating the Prevalence of Sensitive Behavior and Cheating with Dual Design for Direct Questioning and Randomized Response." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59: 723–736. DOI: http://dx.doi.org/10.1111/j.1467-9876.2010.00720.x.

Van der Heijden, P.G.M., G. van Gils, J. Bouts, and J.J. Hox. 2000. "A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning: Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit." *Sociological Methods & Research* 28: 505–537. DOI: http://dx.doi.org/10.1177/0049124100028004005.

Warner, S.L. 1965. "Randomized-Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60: 63–69. DOI: http://dx.doi.org/10.1080/01621459.1965.10480775.

Weissman, A.N., R.A. Steer, and D.S. Lipton. 1986. "Estimating Illicit Drug Use Through Telephone Interviews and the Randomized Response Technique." *Drug and Alcohol Dependence* 18: 225–233. DOI: http://dx.doi.org/10.1016/0376-8716(86)90054-2.

Wolter, F. 2012. Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik. Springer VS.

Wolter, F. and P. Preisendörfer. 2013. "Asking Sensitive Questions: An Evaluation of the Randomized Response Technique versus Direct Questioning Using Individual Validation Data." *Sociological Methods & Research* 42: 321–353. DOI: http://dx.doi.org/10.1177/0049124113500474.

Yu, J.-W., G.L. Tian, and M.L. Tang. 2008. "Two New Models for Survey Sampling With Sensitive Characteristic: Design and Analysis." *Metrika* 67: 251–263. DOI: http://dx.doi.org/10.1007/s00184-007-0131-x.

Received August 2013
Revised October 2014
Accepted October 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 61–75, http://dx.doi.org/10.1515/JOS-2015-0003

Linear Regression Diagnostics in Cluster Samples

Jianzhu Li[1] *and Richard Valliant*[2]

An extensive set of diagnostics for linear regression models has been developed to handle nonsurvey data. The models and the sampling plans used for finite populations often entail stratification, clustering, and survey weights, which renders many of the standard diagnostics inappropriate. In this article we adapt some influence diagnostics that have been formulated for ordinary or weighted least squares for use with stratified, clustered survey data. The statistics considered here include DFBETAS, DFFITS, and Cook's D. The differences in the performance of ordinary least squares and survey-weighted diagnostics are compared using complex survey data where the values of weights, response variables, and covariates vary substantially.

Key words: Cook's D; DFBETAS; DFFITS; influence; model fitting; outlier; residuals.

1. Introduction

Linear regression models and estimators are often applied to analyze complex survey data using the pseudo maximum likelihood (PML) method (e.g., Binder 1983; Skinner et al. 1989).

A sample is considered to be informative when an unweighted model fitted to the sample data is different from the model fitted to the full population (Chambers and Skinner 2003). In such a case, using survey weights in PML estimation accounts for the informativeness. Using the sample weights in the regression estimator not only allows the analysts to account for the design features which govern the data collection process, but also provides a limited type of robustness to model misspecification (Pfeffermann and Holmes 1985; DuMouchel and Duncan 1983; Kott 1991). The sandwich estimator, the Taylor Series linearization estimator (Binder 1983; Fuller 2002), or some type of replication estimator (Wolter 2007) is often employed to obtain both design- and model-consistent variance estimators for the regression parameters. The analyses in this article cover the case in which survey weights are used in regression analysis. If the design is actually noninformative, the diagnostics developed here still apply even though the weights could, in principle, be omitted from model estimation.

Limited attention has been given to diagnosing the adequacy of working models and, more specifically, to detecting outlying and influential observations for regressions using

[1] Westat, 1600 Research Boulevard, Rockville MD 20850, USA. Email: JaneLi@westat.com
[2] Universities of Michigan and Maryland, 1218 Lefrak Hall, College Park MD 20742, USA. Email: rvallian@umd.edu

Acknowledgments: This article is based upon work partially supported by the National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

complex survey data. Different threads of research cover locating and trimming extreme sample weights (Potter 1988, 1990), controlling the effect of outliers on the estimation of descriptive population statistics, and constructing outlier-robust estimation techniques (Chambers et al. 1993; Chambers 1996; Zaslavsky et al. 2001). Henry and Valliant (2012) review much of this literature. Diagnostics for regression models fitted from survey data are a more recent development. Korn and Graubard (1999) and Elliott (2007) introduced techniques for the evaluation of the quality of regressions on complex survey data. Li and Valliant (2009, 2011a, 2011b) examined leverages and methods of identifying influential single observations and groups of observations in single-stage samples. Liao and Valliant (2012a, 2012b) looked at condition indexes and variance inflation factors for linear regressions. In this article we will extend the work of Li and Valliant (2011a) for single-stage samples to samples that use stratification and clustering. We adapt the standard diagnostics – DFBETAS, DFFFITS, and Cook's D – to linear regression models fitted to clustered survey data.

Section 2 specifies the sample design we study, the model that will be used, and a variance estimator that is useful when developing diagnostics. Section 3 presents some diagnostics for identifying single observations that may be influential in fitting a model. Residuals, DFBETAS, DFFITS, and Cook's D are adapted for models fit using stratified, clustered data. In the fourth section, the new diagnostics are illustrated using a data set taken from a large U.S. household survey. Section 5 forms the conclusion.

2. Model Specification and Variance Estimation

To formulate regression diagnostics for clustered survey data, models will be used. Suppose the population contains $h = 1, \ldots, H$ strata, $i = 1, \ldots, N_h$ clusters in stratum h, and $k = 1, \ldots, M_i$ units in cluster hi. A two-stage stratified sample of units is selected with n_h clusters or primary sampling units (PSUs) sampled at the first stage in stratum h with replacement (although without-replacement is more common in practice, a with-replacement formulation has the advantage of producing simpler design-based variance formulas that are more informative for the analyses in this article). The total number of sample clusters is $n = \sum_{h=1}^{H} n_h$. Let m_{hi} be the number of sampled units in the (hi)th cluster, $m = \sum_{h=1}^{H} \sum_{i \in s_h} m_{hi}$, with s_h being the sample of clusters in stratum h, and w_{hik} be the sample weight of the kth unit in the (hi)th cluster. The average number of sample units per sample cluster is $\bar{m} = m/n$. Suppose that $\mathbf{x}_{hik}$ is a p-vector of explanatory variables for unit k in cluster hi and that a variable Y_{hik} collected in the survey follows the linear model:

$$
\begin{aligned}
Y_{hik} &= \mathbf{x}_{hik}^T \boldsymbol{\beta} + \varepsilon_{hik} \\
Cov_M(\varepsilon_{hik}, \varepsilon_{h'i'k'}) &= \begin{cases} \sigma^2 & h = h', i = i', k = k' \\ \sigma^2 \rho & h = h', i = i', k \neq k' \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{1}
$$

This model posits that all units have a common variance and the intracluster correlation, ρ, is the same for all clusters. Units in different clusters are uncorrelated. In practice, ρ is

usually positive and can be estimated using analysis of variance (ANOVA) or related methods. The survey-weighted (SW) estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^{H} \sum_{i \in s_h} \sum_{k \in s_{hi}} \mathbf{A}^{-1} \mathbf{x}_{hik} w_{hik} Y_{hik} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi}$$

with s_{hi} being the sample of units from cluster hi, and

$\mathbf{X}_{hi}$ = the $m_{hi} \times p$ matrix of the $\mathbf{x}_{hik}$ for the m_{hi} sample units in cluster hi;
$\mathbf{W}_{hi}$= the $m_{hi} \times m_{hi}$ diagonal matrix of survey weights for sample units in sample cluster hi;
$\mathbf{Y}_{hi}$ = the m_{hi}-vector of Y_{hik}'s for sample units in cluster hi, and

$$\mathbf{A} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{X}_{hi}.$$

For later use we also define $\mathbf{X}_h^T = (\mathbf{X}_{h1}^T, \ldots, \mathbf{X}_{hn_h}^T)$, $\mathbf{X}^T = (\mathbf{X}_1^T, \ldots, \mathbf{X}_H^T)$, and $\mathbf{W}_h = \text{blkdiag}(\mathbf{W}_{hi})_{i \in s_h}$. Under (1) the model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is

$$\begin{aligned} \text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) &= \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \, \text{var}_M(\mathbf{Y}_{hi}) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \qquad (2) \\ &= \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \left((1-\rho)\sigma^2 \mathbf{I}_{m_{hi}} + \rho\sigma^2 \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \right) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \end{aligned}$$

where $\mathbf{I}_{m_{hi}}$ is the $m_{hi} \times m_{hi}$ identity matrix and $\mathbf{1}_{m_{hi}}$ is a vector of m_{hi} 1s. To test the significance of $\hat{\boldsymbol{\beta}}_{SW}$ or its components, the sandwich estimator in Binder (1983) or the linearization estimator in Fuller (2002) is typically used. Both of these have design-based and model-based justifications. In fact, the sandwich estimator is approximately model unbiased under a model more general than (1), in which the errors are correlated within each cluster but the particular form of the correlation is unspecified (e.g., see Valliant et al. 2000, chap. 9). However, to motivate cutoff values for identifying extremes based on the diagnostics in Section 3, the form of the variance in (2) is useful. Estimates of the components of (2) are needed, and a workable approach is to use purely model-based estimators. To that end, define $\hat{\boldsymbol{\beta}}_{OLS} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{A}_{OLS}^{-1} \mathbf{X}_{hi}^T \mathbf{Y}_{hi}$ with $\mathbf{A}_{OLS} = \sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{X}_{hi}$ to be the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$, and $e_{hik} = Y_{hik} - \mathbf{x}_{hik}^T \hat{\boldsymbol{\beta}}_{OLS}$ to be the residual calculated from the OLS estimator. Using these residuals, define

$$\hat{P} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in s_h} \frac{1}{m_{hi} - 1} \sum_{k \in s_{hi}} (e_{hik} - \bar{e}_{hi})^2$$

$$\hat{Q} = \sum_{h=1}^{H} \sum_{i \in s_h} m_{hi} (\bar{e}_{hi} - \bar{e}_h)^2 \Big/ (n-1)$$

$$\hat{D} = \left(m - \sum_h \sum_{i \in s_h} m_{hi}^2 \Big/ m \right) \Big/ (n-1),$$

where $\bar{e}_{hi} = \sum_{k \in s_{hi}} e_{hik}/m_{hi}$ and $\bar{e}_h = \sum_{i \in s_h} \sum_{k \in s_{hi}} e_{hik} / \sum_{i \in s_h} m_{hi}$. Using $\hat{P}$, $\hat{Q}$, and $\hat{D}$, we can formulate estimators as:

$$\begin{aligned} \widehat{(1-\rho)\sigma^2} &= \hat{P} \\ \widehat{\rho\sigma^2} &= (\hat{Q} - \hat{P})/\hat{D} \end{aligned} \tag{3}$$

These are similar to the estimators in Valliant et al. (2000, sec. 8.3.1) for a common-mean model. Showing that they are model-unbiased for $\rho\sigma^2$ and $(1-\rho)\sigma^2$ is straightforward. Another alternative is to use ANOVA or restricted maximum-likelihood methods in, for instance, SAS® `proc varcomp` or `proc mixed` or Stata® `xtmixed` or the `lmer` function in the R package `lme4` (Bates et al. 2012).

When $\widehat{\rho\sigma^2}$ and $\widehat{(1-\rho)\sigma^2}$ are available, the estimated variance of $\hat{\boldsymbol{\beta}}$ under Model (1) can be constructed as

$$v_M(\hat{\boldsymbol{\beta}}_{SW}) = \sum_h \sum_{s_h} \mathbf{A}^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \left(\widehat{(1-\rho)\sigma^2} \mathbf{I}_{m_{hi}} + \widehat{\rho\sigma^2} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \right) \mathbf{W}_{hi} \mathbf{X}_{hi} \mathbf{A}^{-1} \tag{4}$$

This variance estimator is highly dependent on the working model and is not robust to departures from that model. Because of its nonrobustness, a sandwich or replication estimator is preferred for actually estimating the variance of $\hat{\boldsymbol{\beta}}_{SW}$. However, (4) does have some advantages in determining cutoffs for diagnostics, as described subsequently.

There are alternatives to the estimators of $\rho\sigma^2$ and $(1-\rho)\sigma^2$ in (3). Pfeffermann et al. (1998) proposed the probability-weighted iterative generalized least squares (PWIGLS) estimator to obtain consistent estimates of the population variance parameters σ_U^2 and ρ_U, i.e., the parameters that would be estimated from a census. The PWIGLS estimator, which assumes that the sampling probabilities for both stages π_{hi} and $\pi_{k|hi}$, or equivalently their inverses, w_{hi} and $w_{k|hi}$, are known, is adapted from the standard iterative generalized least squares procedure by analogy with PML. Alternative inflation-type estimators using the two-level sample weights have also been considered (Longford 1995; Graubard and Korn 1996). However, Korn and Graubard (2003) later showed that these estimators can be severely biased when the sampling is informative. They proposed a new set of estimators for variance components that would be approximately unbiased regardless of the sampling design. The limitation of these estimators is that they require knowledge of the second-order inclusion probabilities of the observations. In many surveys, analysts will not know the value of w_{hi}, $w_{k|hi}$, or the joint inclusion probabilities. Consequently, we use the estimators in (3) which are always feasible.

3. Identifying Single Influential Observations

The diagnostic tools presented here are designed to measure the discrepancy in estimated regression coefficients and fitted values, between fitting linear models with and without potentially influential points.

3.1. Residuals

Residuals, which can be used to filter points with outlying Y values, usually are standardized to have unit model variance. For clustered sampling and its corresponding

model (1), we can divide e_{hik} by $\hat{\sigma} = \sqrt{\hat{P} + (\hat{Q} - \hat{P})\hat{D}^{-1}}$; see (3). Generally, the standardized residuals are referred to the standard normal distribution to identify extreme points. If the e_{hik} are not normal, the Gauss inequality (Pukelsheim 1994) is useful for setting a cutoff value.

Gauss Inequality: If the distribution of a random variable X has a single mode at μ_0, then $P\{|X - \mu_0| > r\} \leq 4\tau^2/9r^2$ for all $r \geq \sqrt{4/3}\,\tau$, where $\tau^2 = E\left[(X - \mu_0)^2\right]$.

Suppose that under Model (1), in addition to having a mean of 0, the residuals have a mode of zero. Based on the Gauss Inequality with $r = 2\sigma$, the absolute value of a residual has a probability of about 90% of being less than twice its standard deviation, and with $r = 3\sigma$, it has a probability of about 95% of being less than three times its standard deviation. If we rescale the residuals by a consistent estimate $\hat{\sigma}$ of σ, either $r/\hat{\sigma} = 2$ or 3 can be used to identify outlying residuals, depending on an analyst's preference.

3.2. *DFBETAS*

The standard DFBETAS statistic (Belsley et al. 1980) measures the change in the estimate of $\boldsymbol{\beta}$ when a single unit is removed from the sample. The statistic is also standardized so that it can be referred to a standard normal distribution to determine which values are extreme enough to deserve scrutiny. First, note that (2) can be written as

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \sigma^2 \sum_{h=1}^{H} \sum_{s_h} \mathbf{C}_{hi}\mathbf{R}_{hi}\mathbf{C}_{hi}^T \tag{5}$$

where $\mathbf{R}_{hi} = \left[(1-\rho)\mathbf{I}_{m_{hi}} + \rho \mathbf{1}_{m_{hi}}\mathbf{1}_{m_{hi}}^T\right]$ and $\mathbf{C}_{hi} = \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}$ with (jk)th element $c_{j,hik}$ $(j = 1, \ldots, p; k = 1, \ldots, m_{hi})$. The correlation ρ could be estimated as $\hat{\rho} = \left[1 + \hat{P}\hat{D}/(\hat{Q} - \hat{P})\right]$ or by some other model-based alternative. The variance estimator is then

$$v_M(\hat{\beta}_{SWj}) = \sigma^2 \sum_h \sum_{s_h} \left(c_{j,hi1} \ldots c_{j,him_{hi}}\right) \begin{pmatrix} 1 & & \hat{\rho} \\ & \ddots & \\ \hat{\rho} & & 1 \end{pmatrix} \left(c_{j,hi1} \ldots c_{j,him_{hi}}\right)^T$$

$$= \sigma^2 \sum_h \sum_{s_h} \left(\sum_{k=1}^{m_{hi}} c_{j,hik}^2 + \hat{\rho} \sum_{k \neq k'}^{m_{hi}} c_{j,hik}c_{j,hik'} \right).$$

To measure the difference in each estimated coefficient after the (hik)th unit is deleted, we define $\hat{\boldsymbol{\beta}}_{SW}(hik)$ as the parameter estimate after deleting unit k in cluster hi. The difference between the full sample estimate and the delete-one estimate, $\hat{\boldsymbol{\beta}}_{SW}(hik)$, can be found as

$$DFBETA_{hik} = \hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik) = \frac{\mathbf{A}^{-1}\mathbf{x}_{hik}e_{hik}w_{hik}}{1 - \tilde{h}_{hik,hik}},$$

where $\tilde{h}_{hik,hik} = \mathbf{x}_{hik}^T\mathbf{A}^{-1}\mathbf{x}_{hik}w_{hik}$ is the leverage of the (hik)th unit, which is the kth diagonal element of the matrix $\mathbf{H}_{hii} = \mathbf{X}_{hi}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}$ (see, e.g., Miller 1974;

Valliant et al. 2000, sec. 9.5). The DFBETAS statistic, which is standardized, is constructed as

$$DFBETAS_{hik,j} = \frac{c_{j,hik} e_{hik} / \left(1 - \tilde{h}_{hik,hik}\right)}{\sqrt{\text{var}_M(\hat{\beta}_{SWj})}} \tag{6}$$

$$= \frac{c_{j,hik}}{\sqrt{\sum_{s_h}\left(\sum_{k=1}^{m_i} c_{j,hik}^2 + \rho \sum_{k \neq l}^{m_i} c_{j,hik} c_{j,hil}\right)}} \cdot \frac{e_{hik}}{\sigma} \cdot \frac{1}{1 - \tilde{h}_{hik,hik}}.$$

Note that for actual calculations, a more robust sandwich or replication estimator of $\text{var}_M\left(\hat{\beta}_{SWj}\right)$ would be used in the denominator of (6). Using the diagonal element of (5) in the denominator of $DFBETAS_{hik,j}$ allows us to motivate a heuristic cutoff for identifying extremes.

In order to define a cutoff, some simplifications are needed. If the population and sample sizes from each cluster are bounded by $\bar{M}$ and $\bar{m}$, then $w_{hik} = O\left(N/n\right)$. If the xs are bounded, $\mathbf{C}_{hi} = O\left(n^{-1}\right)$ elementwise and the first term of (6) has order $n^{-1/2}$. Under the same conditions, $\tilde{h}_{hik,hik} = O\left(n^{-1}\right)$, and a rough cutoff after applying the Gauss inequality to e_{hik} would be $2/\sqrt{n}$ or $3/\sqrt{n}$.

A slightly more fine-tuned cutoff is obtained as follows. Following the developments in Scott and Holt (1982) as extended by Liao and Valliant (2012b), the model variance of $\hat{\boldsymbol{\beta}}_{SW}$ can be written as

$$\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right) = \sigma^2 \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{G}$$

where $\mathbf{G} = \left[\sum_{h=1}^{H} \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi}\right] \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1}$. The matrix $\mathbf{G}$ is a generalized design effect that measures the factor by which the model variance differs from that of weighted least squares when all units are uncorrelated. Under Model (1), we have

$$\sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} = \sigma^2 \left[(1 - \rho) \mathbf{X}_h^T \mathbf{W}_h^2 \mathbf{X}_h + \rho \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{W}_{hi}^2 \mathbf{X}_{Bhi}\right].$$

where $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}$ with $\mathbf{1}_{m_{hi}}$ being a vector of m_{hi} 1s. If the sample is self-weighting so that $w_{hik} \equiv w$, then under Model (1) $\mathbf{G}$ can be written as

$$\mathbf{G} = w\sigma^2 \left[\mathbf{I}_p + (\mathbf{M} - \mathbf{I}_p)\rho\right]$$

where $\mathbf{M} = \left(\sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{X}_{Bhi}\right) (\mathbf{X}^T \mathbf{X})^{-1}$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix. If we assume that the sample size within every cluster is $m_{hi} = \bar{m}$ and that the vector of covariates for every element in cluster hi is the same, $\mathbf{x}_{hik} = \bar{\mathbf{x}}_{hi}$, with some algebra it

follows that

$$\sum_{i\in s_h} m_{hi}\mathbf{X}_{Bhi}^T\mathbf{X}_{Bhi} = \bar{m}\sum_h\sum_{s_h}\bar{\mathbf{x}}_{hi}\bar{\mathbf{x}}_{hi}^T$$

$$\mathbf{X}^T\mathbf{X} = \sum_h\sum_{s_h}\bar{\mathbf{x}}_{hi}\bar{\mathbf{x}}_{hi}^T.$$

Using these results, $\mathbf{M}$ reduces to $\bar{m}\mathbf{I}_p$. In these special circumstances, the model variance of the survey-weighted least squares estimator is

$$\text{var}_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\left[\mathbf{I}_p + \rho\times diag(\bar{m}-1)\mathbf{I}_p\right].$$

The model variance of the jth coefficient of $\hat{\boldsymbol{\beta}}_{SW}$, which is needed for $DFBETAS_{hik,j}$, is then

$$\text{var}_M(\hat{\beta}_{SWj}) = \sigma^2(\mathbf{X}^T\mathbf{X})_{jj}^{-1}[1+(\bar{m}-1)\rho]$$

where $(\mathbf{X}^T\mathbf{X})_{jj}^{-1}$ denotes the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Assuming the xs are all bounded, the order of magnitude of each element of $(\mathbf{X}^T\mathbf{X})^{-1}$ is n^{-1}. Thus $\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SWj}\right) = O\left(n^{-1}\right)\left[1+(\bar{m}-1)\rho\right]$. Using $c_{j,hik} = O\left(n^{-1}\right)$, the first term in (6) is $c_{j,hik}\Big/\sqrt{\text{var}_M(\hat{\beta}_j)} \approx \left\{O(n)\left[1+(\bar{m}-1)\rho\right]\right\}^{-1/2}$. As a result, a somewhat more refined cutoff value for $DFBETAS_{ik,j}$ is $2\Big/\sqrt{n\left[1+(\bar{m}-1)\rho\right]}$ or $3\Big/\sqrt{n\left[1+(\bar{m}-1)\rho\right]}$.

3.3. DFFITS

Multiplying the DFBETA statistic by the $\mathbf{x}_{hik}^T$ vector, we obtain the measure of change in the (*hik)*th fitted values due to the deletion of the *(hik)*th observation,

$$DFFIT_{hik} = \hat{Y}_{hik} - \hat{Y}_{hik}(hik) = \frac{\tilde{h}_{hik,hik}e_{hik}}{1-\tilde{h}_{hik,hik}}.$$

The variance of the predicted value is

$$\text{var}_M\left(\hat{Y}_{hik}\right) = \mathbf{x}_{hik}^T\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right)\mathbf{x}_{hik}$$

$$= \sigma^2\sum_{i'\in s}\left(\sum_{k'=1}^{m_{hi'}}\tilde{h}_{hik,hi'k'}^2 + \rho\sum_{k''\neq k'}^{m_{hi'}}\tilde{h}_{hik,hi'k'}\tilde{h}_{hik,hi'k''}\right).$$

The DFFITS statistic is formulated as

$$DFFITS_{hik} = \frac{\tilde{h}_{hik,hik}e_{hik}/\left(1-\tilde{h}_{hik,hik}\right)}{\sqrt{\text{var}_M\left(\hat{Y}_{hik}\right)}}$$

We can make approximations analogous to the ones used for DFBETAS in order to justify a cutoff for DFFITS. Based on (7) for the special case of $m_{hi} = \bar{m}$ and $\mathbf{x}_{hik} = \bar{\mathbf{x}}_{hi}$, we have $v_M\left(\hat{Y}_{ik}\right) = \mathbf{x}_{ik}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left[\mathbf{I}_p + diag(\bar{m}-1)\rho\right]\mathbf{x}_{ik}$. Each element of $\mathbf{X}^T\mathbf{X}$ is the sum of m elements, and, if each x is bounded, is $O(m)$. The variance $\text{var}_M\left(\hat{Y}_{ik}\right)$ is a sum of

p elements; thus $v_M\left(\hat{Y}_{ik}\right) = O(p/m)\left[1 + (\bar{m} - 1)\rho\right]$. Since the average leverage is p/m, a rough value on $\frac{\tilde{h}_{hik,hik}/\left(1-\tilde{h}_{hik,hik}\right)}{\sqrt{\text{var}_M\left(\hat{Y}_{hik}\right)}}$ is $\frac{p/m}{1-p/m} \Big/ \sqrt{\frac{p}{m}\left[1 + (\bar{m} - 1)\rho\right]} = \sqrt{p/\left\{n\bar{m}\left[1 + (\bar{m} - 1)\rho\right]\right\}}$, assuming that the number of sample units, m, is much larger than the number of regressors, p. Thus a heuristic cutoff for the DFFITS statistic is $k\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$ with k being 2 or 3.

3.4. Modified Cook's Distance

Under the working Model (1), a quadratic statistic that measures the effect on the entire $\hat{\boldsymbol{\beta}}_{SW}$ vector of dropping the kth element in cluster hi can be constructed as

$$ED_{hik} = \left[\hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik)\right]^T \left[\text{var}\left(\hat{\boldsymbol{\beta}}_{SW}\right)\right]^{-1} \left[\hat{\boldsymbol{\beta}}_{SW} - \hat{\boldsymbol{\beta}}_{SW}(hik)\right]$$

where $\hat{\boldsymbol{\beta}}_{SW}(hik)$ is the parameter estimate after deleting unit k in cluster hi and $\text{var}\left(\hat{\boldsymbol{\beta}}_{SW}\right)$ is any of the variance estimators discussed in Section 1. To determine a heuristic cutoff value for ED_{ik}, we use the model variance $\text{var}_M\left(\hat{\boldsymbol{\beta}}_{SW}\right)$ under (1) and write the statistic as

$$ED_{hik} = \left(\frac{e_{hik}}{\sigma}\right)^2 \frac{1}{\left(1 - \tilde{h}_{hik,hik}\right)^2} w_{hik}\mathbf{x}_{hik}^T \left[\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{x}_{hik} w_{hik}$$

where the matrix $\mathbf{R}$ is block diagonal with 1 on the diagonal and ρ off the diagonal in each block (cluster); the dimension of block hi is $m_{hi} \times m_{hi}$. If the number of units within each sampled PSU, m_{hi}, is bounded, $w_{hik}\mathbf{x}_{hik}^T \left[\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{x}_{hik} w_{hik} = O\left(n^{-1}\right)$, and using similar reasoning to that employed in Subsections 3.1 and 3.2, we arrive at a rough value for ED_{hik} of $p\left[n\bar{m}\left(1 + \hat{\rho}(\bar{m} - 1)\right)\right]^{-1}$. Therefore, in the clustered sampling case we can compare $\sqrt{ED_{hik}}$ with the cutoff value $2\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$ or $3\sqrt{p/n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]}$. A more convenient form is found by standardizing ED_{hik} and taking its square root. Based on the classic Cook's Distance, we term this the Modified Cook's distance:

$$MD_{hik} = \sqrt{\left\{n\bar{m}\left[1 + \hat{\rho}(\bar{m} - 1)\right]\right\} ED_{hik}/p}$$

and compare MD_{hik} to 2 or 3.

Table 1. Quantiles of variables in NHANES regression of systolic blood pressure on age, BMI, and blood lead

	Quantiles				
Variables	0%	25%	50%	75%	100%
Systolic BP	82	102	108	114	146
Age	20	22	24	27	29
BMI	14.42	22.84	26.43	31.62	61.68
Log(Lead+1)	0.18	0.47	0.64	0.83	3.75
Survey Weight	698.39	3,576.69	11,467.06	31,094.18	103,831.17

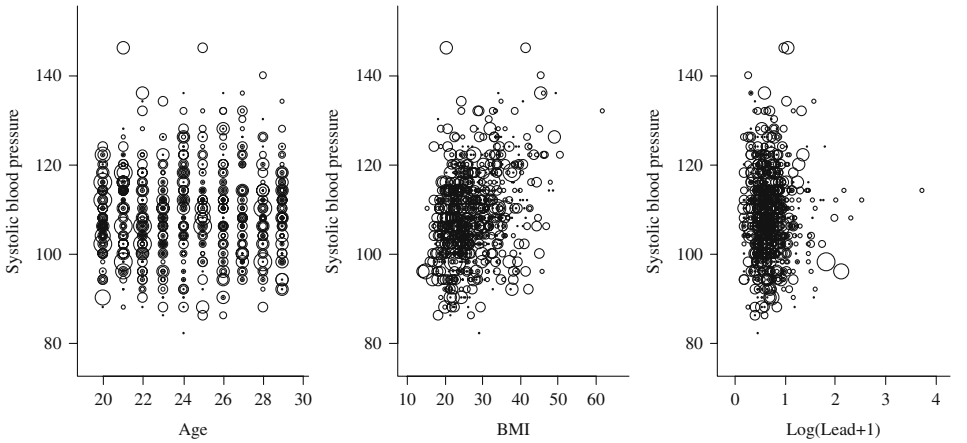

Fig. 1. Bubble plots of systolic blood pressure versus three auxiliary variables for NHANES data. The areas of the bubbles are proportional to sample weights

4. Case Study: NHANES

In this section, we examine a regression of systolic blood pressure on the logarithm of blood lead level, age, and body mass index using a subset from the National Health and Nutrition Examination Survey (NHANES) 1999-2002. The subset used in this study has a sample size of 810, consisting of Mexican-American females aged 20 to 29. This sample does not have very skewed $\mathbf{Y}$ and $\mathbf{X}$ values, but involves clustering and stratification in the sampling design with a set of large and greatly varying sample weights. There are $n = 57$ PSUs nested in $H = 28$ strata, all but one of the strata having 2 PSUs. The average cluster size $\bar{m}$ is 14.21 persons. When applied to a clustered data set, the variance estimators in the survey-weighted diagnostic statistics need to take the design into account and the cutoffs

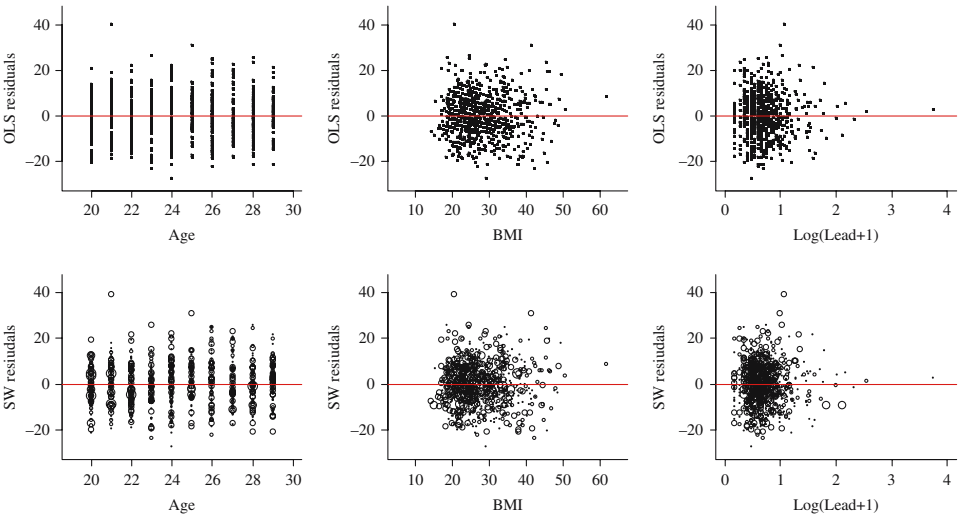

Fig. 2. OLS and SW residuals versus three auxiliary variables for NHANES data. Horizontal reference lines are drawn at zero

Table 2. OLS and SW parameter estimates from NHANES regression

	OLS Estimation			SW Estimation		
Independent Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	94.91***	3.11	30.55	99.79***	4.72	21.16
Age	0.02	0.11	0.14	−0.15	0.17	−0.87
BMI	0.45***	0.05	9.23	0.44***	0.07	5.88
Log(Lead+1)	1.03	0.99	1.04	0.89	1.28	0.70

*** Significant at level 0.001

for some of the statistics contain an estimate of ρ, which in Model (1) describes the correlation between the observations within the same cluster. The illustrative calculations in this study do not account for the fact that Mexican-American females are a domain within the full population whose sample size is random. This will tend to make SW variance estimates smaller than they would be if the domain feature was accounted for.

Table 1 gives the quantile values of the variables and sample weights used in the regression. Besides demonstrating the skewness and large range of sample weights, the table also shows that the distributions of BMI and the logarithm of the blood lead are skewed to the right. Since the minimum of the originally measured blood lead level is as small as 1, we added 1 to blood lead level before taking the logarithm to generate positive transformed values. (Adding 1 is often done to avoid taking the log of zero; this step was not strictly necessary here.) Note that using the untransformed value of blood lead would have resulted in more extreme **X** values. However, this type of modeling has previously been done using the log transformation (see Korn and Graubard 1999), and we follow that precedent here. Figures 1 and 2 respectively display plots of systolic blood pressure and residuals versus the three auxiliary variables. Table 2 reports the parameter estimates of the regressions with and without weights. The SW estimators produced slightly larger intercept and slightly smaller slope of BMI than the OLS ones. Both methods agree that

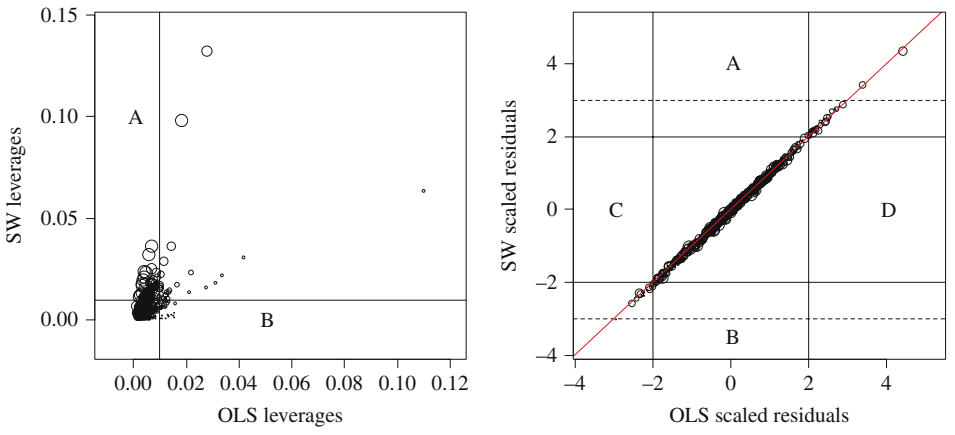

Fig. 3. Leverage and residual plots for NHANES data. In left-hand panel, A = points identified only by SW diagnostics; B = points identified only by OLS diagnostics; vertical and horizontal reference lines are drawn at $2p/n\tilde{m}$. In right-hand panel, A,B = points identified by SW but not OLS. C,D = points identified by OLS but not SW

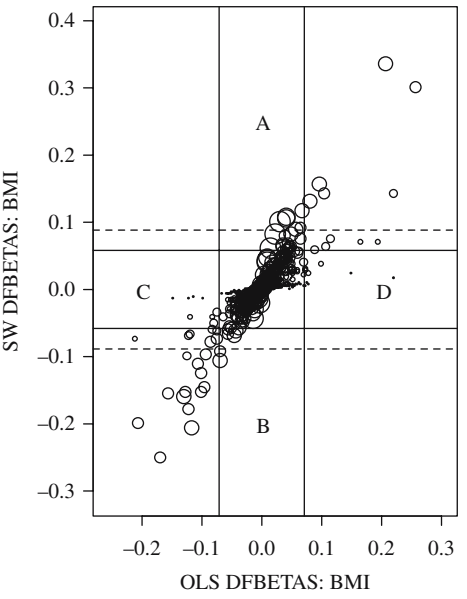

Fig. 4. DFBETAS Plot of BMI for NHANES Data. A,B = points identified by SW but not OLS. C,D = points identified by OLS but not SW

age and blood lead do not have significant effects in determining the systolic blood pressure. Therefore, in the following diagnostic analysis, we will only focus on the changes in the estimated coefficient of BMI.

For comparison, we applied both the OLS and the new SW diagnostic statistics, including leverages, residuals, DFBETAS, DFFITS, and modified Cook's distance, to the regression estimation. Since the sample weights were not separately provided at cluster level and at unit level, the parameters ρ and σ^2 in Model (1) were estimated using purely model-based estimators. Utilizing the `VARCOMP` procedure in SAS, we obtained $\hat{\rho} = 0.033$ and $\hat{\sigma}^2 = 82.09$. The design effect was estimated as $\sqrt{1 + \hat{\rho}(\bar{m} - 1)} = 1.2$. For the

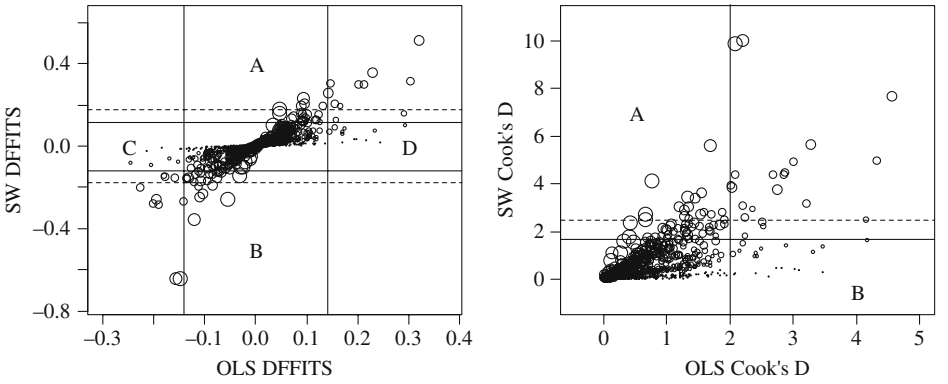

Fig. 5. DFFITS plot and modified Cook's distance plot for NHANES data. In left-hand panel A,B = points identified by SW but not OLS; C,D = points identified by OLS but not SW. In right-hand panel A = points identified by SW but not OLS; B = points identified by OLS but not SW

Table 3. Number of outliers identified and associated weight ranges for NHANES data

	Outliers identified by OLS only		Outliers identified by SW only	
Diagnostic statistics	Counts	Weight range	Counts	Weight range
Leverage	24	(875.5, 13,085.8)	85	(16,929.6, 103,831.2)
Residual	1	(2,730.1, 2,730.1)	8	(1,791.1, 36,955.3)
DFBETAS(BMI)	25	(1,773.5, 2,3677.5)	12	(32,451.1, 103,831.2)
DFFITS	21	(994.9, 17,366.9)	28	(2,9617.1, 103,831.2)
Modified Cook's D	21	(994.9, 17,366.9)	35	(21,194.0 103,831.2)

SW diagnostics, a strict criterion, 2, was used to construct cutoffs. For example, the cutoff of DFBETAS is $2\Big/\sqrt{n\bar{m}\left[1+\hat{\rho}(\bar{m}-1)\right]}$. The solid reference lines in the subsequent figures were drawn at the cutoff values of 2; dotted reference lines using the looser criterion of 3 are also drawn in the same graphs.

Figures 3 through 5 display the comparisons between the OLS and the SW diagnostic statistics. The range of the weights in the NHANES data set is extremely wide, with a minimum of 698.39 and a maximum of 103,831.17. Hence the SW diagnostics tend to identify more influential observations with large weights, whereas the OLS diagnostics tend to detect more points with small weights. The leverage plot (Figures 3), DFBETAS plot (Figure 4), and the modified Cook's distance plot (Figure 5) clearly show that the "identified by SW only" areas contain many big bubbles, but the "identified by OLS only" areas are filled with small dots. The residual plot is an exception in which the OLS and the SW residuals are very similar. This is mainly because none of the **Y** and **X** values in the data set are extremely outlying.

Table 3 numerically reports the weight discrepancies between the observations uniquely identified by either OLS or SW diagnostics. The leverage and modified Cook's distance are more sensitive to extreme sample weights compared to other diagnostic statistics. They tend to detect more influential points for survey data than the OLS approaches. Analysts may want to consider raising the cutoff values for these statistics in order not to overidentify influential points.

Table 4. Estimated slopes of BMI from full sample and reduced samples by different diagnostic approaches for NHANES data

	OLS estimation			SW estimation		
	BMI	SE	*t*	BMI	SE	*t*
Full sample	0.45***	0.05	9.23	0.44***	0.07	5.88
Leverages	0.39***	0.06	6.86	0.43***	0.08	5.23
Residuals	0.47***	0.04	10.50	0.47***	0.06	8.19
DFBETAS (BMI)	0.49***	0.05	9.51	0.46***	0.05	8.83
DFFITS	0.47***	0.05	9.76	0.45***	0.05	8.51
Modified Cook's D	0.47***	0.05	9.76	0.44***	0.05	8.74

*** Significant at level 0.001

The parameter estimates after outliers were removed are listed in Table 4. The difference between the OLS and SW estimates and the two diagnostic schemes is trivial. The removal of observations with large DFBETAS of BMI causes the largest change in the estimated slope of BMI. The SW estimates seem to be less affected by the removal of influential points than the OLS ones. Unlike the SMHO data analyzed in Li and Valliant (2011a), the NHANES data set does not contain many obviously extreme points, and outlying Y values can be large or small relative to other points. Hence the deletion of the identified outliers does not move the regression line dramatically.

5. Conclusion

By incorporating survey weights and design features, we constructed survey-weighted diagnostic statistics for clustered samples that are extensions of the conventional OLS diagnostics. Survey-weighted diagnostics may identify different points than OLS diagnostics as influential. An observation with moderate Y and $\mathbf{x}$ values may not be identified as influential by OLS approaches, but may be recognized as influential by SW methods if it is assigned an extreme sample weight. The diagnostics can serve as a guide to which points may be unusual. However, a diligent analyst should examine these points in detail to decide whether they are data entry errors, legitimate values that do not follow a core model, or can be explained in some other way, such as having extreme weights.

The techniques based on single-case deletion presented here may not function effectively when some outliers mask the effects of others. The modified forward search method (Atkinson and Riani 2000, 2004; Li and Valliant 2011b) is a partial solution to this problem since it can successfully identify an influential group of points whose members are not influential when examined singly.

A final caveat to the use of the diagnostics studied here is that some points may appear to be influential because the regression model itself is misspecified. Deleting them would be a mistake if the ability is lost to recognize that the model should be respecified, for example, as quadratic. Thus good practice will require using a combination of residuals and the other diagnostics studied here.

6. References

Atkinson, A.C., and M. Riani. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.

Atkinson, A.C., and M. Riani. 2004. "The Forward Search and Data Visualization." *Computational Statistics* 19: 29–54.

Bates, D., M. Maechler, B. Bolker and S. Walker. 2014. "*lme4: Linear Mixed-Effects Models Using Eiqen and S4. R package version 1.1-7*." Available at: http://CRAN.R-project.org/package= lme4 (accessed February 2, 2015).

Belsley, D.A., R. E. Kuh, and R. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Binder, D.A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–292. DOI: http://dx.doi.org/10.2307/1402588

Chambers, R.L., A.H. Dorfman, and T.E. Wehrly. 1993. "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration." *Journal of the American Statistical Association* 88: 268–277. DOI: http://dx.doi.org/10.1080/01621459.1993.10594319

Chambers, R.L. 1996. "Robust Case-Weighting for Multipurpose Establishment Surveys." *Journal of Official Statistics* 12: 3–32.

Chambers, R.L., and C.J. Skinner. 2003. *Analysis of Survey Data*. New York: John Wiley.

DuMouchel, W.H., and G.J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples." *Journal of the American Statistical Association* 78: 535–543. DOI: http://dx.doi.org/10.1080/01621459.1983.10478006

Elliott, M. 2007. "Bayesian Weight Trimming for Generalized Linear Regression Models." *Survey Methodology* 33: 23–34.

Fuller, W.A. 2002. "Regression Estimation for Survey Samples." *Survey Methodology* 28: 5–23.

Graubard, B.I., and E.L. Korn. 1996. "Modelling the Sampling Design in the Analysis of Health Surveys." *Statistical Methods in Medical Research* 5: 263–281. DOI: http://dx.doi.org/10.1177/096228029600500304

Henry, K.A., and R. Valliant. 2012. "Methods for Adjusting Survey Weights When Estimating a Total." In *Proceedings of the Federal Committee on Statistical Methodology*, January 10–12. Washington, DC. Available at: http://fcsm.sites.usa.gov/files/2014/05/Henry_2012FCSM_V-A.pdf (accessed February 2, 2015)

Korn, E.L., and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Korn, E.L., and B.I. Graubard. 2003. "Estimating Variance Components by Using Survey Data." *Journal of Royal Statistical Society B* 65: 175–190. Part 1. DOI: http://dx.doi.org/10.1111/1467-9868.00379

Kott, P.S. 1991. "A Model-Based Look at Linear Regression with Survey Data." *American Statistician* 45: 107–112. DOI: http://dx.doi.org/10.1080/00031305.1991.10475779.

Li, J., and R. Valliant. 2009. "Survey Weighted Hat Matrix and Leverages." *Survey Methodology* 35: 15–24.

Li, J., and R. Valliant. 2011a. "Linear Regression Influence Diagnostics for Unclustered Survey Data." *Journal of Official Statistics* 27: 99–119.

Li, J., and R. Valliant. 2011b. "Detecting Groups of Influential Observations in Linear Regression using Survey Data—Adapting the Forward Search Method." *Pakistan Journal of Statistics* 27: 507–528.

Liao, D., and R. Valliant. 2012a. "Variance Inflation Factors in the Analysis of Complex Survey Data." *Survey Methodology* 38: 53–62.

Liao, D., and R. Valliant. 2012b. "Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data." *Survey Methodology* 38: 189–202.

Longford, N.T. 1995. *Models for Uncertainty in Educational Testing*. New York: Springer-Verlag.

Miller, R.G., Jr. 1974. "An Unbalanced Jackknife." *The Annals of Statistics* 2: 880–891.

Pfeffermann, D., and D.J. Holmes. 1985. "Robustness Considerations in the Choice of Method of Inference for the Regression Analysis of Survey Data." *Journal of the Royal Statistical Society A* 148: 268–278. DOI: http://dx.doi.org/10.2307/2981971

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society B* 60: 23–40. DOI: http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00106/abstract

Potter, F.A. 1988. "Survey of Procedures to Control Extreme Sampling Weights." In *Proceedings of the Section on Survey Research Methods: American Statistical Association*, 453–458. Available at: http://www.amstat.org/sections/SRMS/proceedings/.

Potter, F.A. 1990. "Study of Procedures to Identify and Trim Extreme Sample Weights." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, 225–230. Available at: http://www/amstat.org/sections/SMRM/proceedings/.

Pukelsheim, F. 1994. "The Three Sigma Rule." *The American Statistician* 48: 88–91.

Scott, A.J., and D. Holt. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of the American Statistical Association* 77: 848–854.

Skinner, C.J., D. Holt, and T.M.F. Smith (eds.). 1989. *Analysis of Complex Surveys*. New York: Wiley.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Wolter, K. 2007. *Introduction to Variance Estimation*. New York: Springer.

Zaslavsky, A., N. Schenker, and T. Belin. 2001. "Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey." *Journal of the American Statistical Association* 96: 858–869.

Received August 2013
Revised May 2014
Accepted September 2014

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 77–100, http://dx.doi.org/10.1515/JOS-2015-0004

Ratio Edits Based on Statistical Tolerance Intervals

Derek S. Young[1] *and Thomas Mathew*[2]

The role of statistical tolerance intervals for developing ratio edit tolerances in a parametric setup is investigated. The performance of the methodology is assessed for the normal and Weibull distributions. The numerical results show that in terms of Type I and Type II errors, statistical tolerance intervals exhibit better performance compared to other ratio edit procedures available in the literature. The methodology is illustrated using 2010 and 2011 data from the Annual Survey of Manufacturers.

Key words: Outliers; resistant; robust; tolerance limits; trimming; Winsorization.

1. Introduction

Ratio edit tolerances are bounds used for identifying errors in the data obtained by Economic Census Programs so that they can be flagged for further review. The tolerances represent upper and lower bounds on the ratio of two highly correlated items and are used for outlier detection; that is, to identify units that are inconsistent with the rest of the data. Some texts dedicated to the general topic of outlier detection include Barnett and Lewis (1994), Rousseeuw and Leroy (2003), and Aggarwal (2013). A number of outlier detection methods are also available in the literature and can be used for developing ratio edit tolerances; we refer to Thompson and Sigman (1999) and Rais (2008) for a review and comparison of these methods as they apply to the ratio edit problem. Thompson and Sigman (1999) compared different methods for generating ratio edit tolerances, which focused on "Type I" and "Type II" errors. A Type I error flags a ratio value as inconsistent or wrong when it is not so. A Type II error flags an inconsistent ratio as consistent or correct. Thompson and Sigman (1999) recommended a stepwise approach for developing ratio edit tolerances, while Thompson and Adeshiyan (2003) discussed the effects of ratio edit and imputation procedures on data quality for the 1997 Economic Census. Both articles also emphasized the importance of incorporating subject-matter expertise when developing the ratio edits.

[1] Department of Statistics, University of Kentucky, 725 Rose Street, Lexington, KY 40536, USA. Email: derek.young@uky.edu
[2] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A. Email: thomas.mathew@census.gov
Acknowledgments: This research was conducted while the first author was a Research Mathematical Statistician in the Center for Statistical Research and Methodology, U.S. Census Bureau. The authors wish to thank Jenny Thompson and Eric Slud of the U.S. Census Bureau, three anonymous referees, and the Associate Editor for their numerous helpful comments on this paper. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

The issue of outliers or large data values in surveys has been addressed in the literature. Tambay (1988) presents an empirical study comparing various methodologies for identifying level outliers and/or trend outliers in subannual economic surveys. Latouche and Berthelot (1992) focus on respondent follow-ups to units that may have an important effect on statistical estimates. The authors present and compare three score functions as a way to identify suspicious units according to their potential effect on the estimates. Kokic and Bell (1994) discuss the setting where a number of unusually large observations fall in the survey sample, which may grossly overestimate population totals. They proceed to specify a cutoff criterion so that an optimal level can be found for Winsorizing the data. As discussed in Rivest and Hidiroglou (2004), Winsorization is widely used to curb the effect of outliers when computing survey estimates. Winsorized estimates have a downward bias and smaller variance relative to their non-Winsorized analogues. When aggregating survey estimates, these effects result in larger biases and less precision than standard aggregated estimates. Hence, Rivest and Hidiroglou (2004) propose using a "corrected" Winsorized estimate.

While not investigated here, we note a few other novel outlier detection methods that could be investigated for performing ratio edits. Hido et al. (2011) present an approach to identify outliers in a test dataset based on a training dataset comprised solely of inliers, which is accomplished by using the ratio of the two dataset densities as an outlier score. Yuen and Mu (2012) use a Bayesian linear regression setup to compute probabilities that an observation is an outlier. Finally, Chawla and Gionis (2013) present a generalization to the k-means algorithm as a way to simultaneously cluster and discover outliers in a dataset.

The purpose of our investigation is to examine the role of statistical tolerance intervals in the process of developing ratio edit tolerances. A statistical tolerance interval provides bounds that will capture a specified proportion or more of a sampled population with a given confidence level; we refer to the book by Krishnamoorthy and Mathew (2009) for a detailed discussion of the topic. Since ratio edit tolerances provide a range for the acceptable ratios, a statistical tolerance interval can do the same provided that such an interval is constructed using the good ratios; that is, using the data after deleting the ratios that are inconsistent or problematic. An advantage of using a statistical tolerance interval is that such an interval, by construction, controls the Type I error at a specified level, similar to what is done in hypothesis testing. The Type II error performance can then be studied and compared with other ratio edit tolerance intervals available in the literature, as described in Thompson and Sigman (1999).

Our approach consists of computing statistical tolerance intervals based on the "good" part of the data; that is, after trimming the data so that potentially bad ratios are excluded from the tolerance interval computation. We have no clear guidance on the percentage of trimming to be done, which should perhaps be done using the input of a subject-matter expert. In the case of a nearly symmetric distribution, we recommend trimming both tails of the distribution, unless there is reason to believe that the contamination is only in one tail. We report numerical results for a two-sided tolerance interval for the case of a normal distribution, computed after trimming both tails. Type I and Type II error probabilities are reported and compared with the ratio edit tolerances available in the literature. We also report results for a one-sided upper statistical tolerance limit for the case of a Weibull distribution, computed after trimming is done only in the right tail. The overall conclusion

is that the statistical tolerance interval approach has a considerable edge over the available ratio edit tolerances in terms of controlling Type I and Type II error probabilities. Furthermore, for several standard parametric distributions (including the normal and Weibull distributions considered in our work), analytic expressions or accurate approximations are available for the limits that define a statistical tolerance interval. In other words, they are easy to compute and we refer to Krishnamoorthy and Mathew (2009) for further details.

Before describing the methodology for computing a statistical tolerance interval, we want to make a brief comment on the terminology used in this article and in the literature. As already noted, ratio edit tolerances are thresholds used for identifying ratio edit failures and are determined through a wide range of possible outlier detection methods; however, they are not defined or determined using the same criteria that define a statistical tolerance interval. On the other hand, statistical tolerance limits are bounds that capture at least a specified proportion of the sampled population with a given confidence level. Since both notions are traditionally referred to as "tolerance limits," we will make it clear through the context which type of "tolerance" is being discussed.

We begin our discussion with a review of outlier detection methods that are used for ratio edits and then investigate the role of statistical tolerance intervals for the same.

2. Outlier Detection Methods for Ratio Edits

There are numerous procedures for outlier detection in the literature; for example, see the texts by Iglewicz and Hoaglin (1993), Barnett and Lewis (1994), and Rousseeuw and Leroy (2003). The focus of this study is not to provide an exhaustive comparison of those procedures, but rather to compare our approach with the standard methods used in setting ratio edit tolerances. In this section, we discuss three common approaches that have been employed by the U.S. Census Bureau.

2.1. Robust Control Limits

Shewhart (1939) provided the first thorough treatment of control charts as a way to monitor a quality characteristic of a process over time. Control charts (also called Shewhart charts) are a simple, yet powerful way to visualize variability in a process. They can be used to identify shifts in a process or when a process goes out of control, where this latter setting is essentially an outlier detection problem. The outliers are identified by placing control limits on the data. Let μ_T and σ_T denote the mean and standard deviation, respectively, of a statistic of interest $T \equiv T(X)$ for the process being monitored. Then lower and upper control limits are given by $\mu_X - L\sigma_X$ and $\mu_X + L\sigma_X$, respectively. Here, L controls how far one will allow the process to vary from the mean before determining that it has gone "out of control." Typically, we set $L = 3$, which is the 3σ-limit rule of thumb often used for outlier detection. A more contemporary treatment of control chart methodology can be found in Montgomery (2013).

While ratio data is usually not time ordered (even though the ratios themselves may be constructed using the same variable measured at two different time points), we can still apply a similar type of control limit methodology. As discussed in Thompson and Sigman (1999), we can use robust estimates of the population mean and standard deviation to

construct control limits, which in turn will be the ratio edit tolerances. The robust estimates are based on trimming and Winsorizing, which we now describe in more detail for any general univariate setting.

Suppose we have observed data $x_1, \ldots, x_n$ and let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ denote the ordered data. The (symmetric) α-trimmed mean for the data is given by

$$\bar{x}_\alpha = \frac{1}{n - 2\lceil \alpha n \rceil} \sum_{i=\lceil \alpha n \rceil + 1}^{n - \lceil \alpha n \rceil} x_{(i)}, \tag{1}$$

where $\lceil \cdot \rceil$ is the ceiling function and $0 < \alpha < 1$. As noted in Tukey and McLaughlin (1963), the Winsorized variance is a consistent estimator of the variance of (1). The Winsorized variance is given by

$$s^2_{W_\alpha} = \frac{1}{n - 2\lceil \alpha n \rceil} \sum_{i=\lceil \alpha n \rceil + 1}^{n - \lceil \alpha n \rceil} (x_{(i)} - \bar{x}_{W_\alpha})^2, \tag{2}$$

where

$$\bar{x}_{W_\alpha} = \frac{1}{n} \left(\sum_{i=\lceil \alpha n \rceil + 1}^{n - \lceil \alpha n \rceil} x_{(i)} + \lceil \alpha n \rceil \left(x_{(\lceil \alpha n \rceil + 1)} + x_{(n - \lceil \alpha n \rceil)} \right) \right) \tag{3}$$

is the Winsorized mean. It is easy to modify the above formulas to accommodate asymmetric trimming and Winsorizing, which includes one-sided trimming and Winsorizing as special cases. Finally, the interval based on robust control limits is given by

$$(\bar{x}_\alpha - L s_{W_\alpha}, \bar{x}_\alpha + L s_{W_\alpha}). \tag{4}$$

For ratio data, Thompson and Sigman (1999) use $L = 2$ to set a more liberal rule and $L = 3$ to set a more conservative rule regarding the number of cases flagged for review.

Many robust measures of location and scale could be investigated to construct analogues to the robust control limits in Equation (4). For example, one might simply consider the median or an M-estimator for a robust estimate of location, while the median absolute deviation or Gini's mean difference could be used for a robust estimate of scale. These may result in more informative limits for a particular application. However, our focus is on comparing some of the more common methods used in setting ratio edit tolerances (e.g., Equation (4)) with the tolerance interval approach that we discuss in Section 3.

2.2 *Fence-Based Methods*

In exploratory data analysis, the interquartile range (IQR) can be used to identify potential outliers in a univariate dataset. The IQR is a resistant measure of dispersion defined as $Q_3 - Q_1$, where Q_1 and Q_3 are the first and third quartiles, respectively. As discussed in Hoaglin et al. (1986), the resistant rule flags values as outliers if they fall outside the interval

$$(Q_1 - k\text{IQR}, Q_3 + k\text{IQR}), \tag{5}$$

for some non-negative constant k. Thompson and Sigman (1999) studied the use of Equation (5) as a way to set ratio edit tolerances and referred to the above rule as *resistant fences*. They referred to the specific rules of setting the values of k equal to 1.5, 2.0, and 3.0 as inner, middle, and outer fences, respectively. Note that the inner-fences rule is almost always employed when identifying univariate outliers on a boxplot.

Thompson (1999) explored a variation of resistant fences for asymmetric distributions. *Asymmetric fences* are elongated in the direction of the skewness of the distribution. Denoting the median by $\tilde{x}$, the asymmetric-fences method replaces the IQR in Equation (5) with distances from $\tilde{x}$. Specifically, the asymmetric-fences rule flags values as outliers if they fall outside the interval

$$(Q_1 - k^*(\tilde{x} - Q_1), Q_3 + k^*(Q_3 - \tilde{x})). \tag{6}$$

For asymmetric fences, Thompson (1999) refers to values of $k*$ equal to 3.0, 4.0, and 6.0 as inner, middle, and outer fences. Note that these rules are just twice the value of k used for the resistant-fences rule.

2.3 *Hidiroglou-Berthelot Method*

The methodology introduced by Hidiroglou and Berthelot (1986) is a ratio edit procedure that uses a *centering transformation* of the ratios followed by a *magnitude transformation*. Here is a brief description of the procedure.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be observations of the variables of interest and $r_i = x_i/y_i$, $i = 1, \ldots, n$ denote the n ratios to be analyzed. Moreover, let $\tilde{r}$ denote the median of the ratios. Define

$$s_i = \begin{cases} (r_i/\tilde{r}) - 1, & \text{if } r_i \geq \tilde{r} \\ 1 - (\tilde{r}/r_i), & \text{if } r_i < \tilde{r} \end{cases} \tag{7}$$

and

$$e_i = s_i \times (\max\{x_i, y_i\})^U, \tag{8}$$

where $0 \leq U \leq 1$. As noted by Hidiroglou and Berthelot (1986), the quantity U "provides control on the importance associated with the magnitude of the data"; see also Thompson (2007). The values $U = 0.30$ and $U = 0.50$ are recommended in Belcher (2003), Sigman (2002), and Thompson (2007).

Next, let e_{Q_1}, $\tilde{e}$ and e_{Q_3} denote, respectively, the first quartile, the median, and the third quartile of the e_i's. Now define $d_{Q_1} = \max\{\tilde{e} - e_{Q_1}, |A\tilde{e}|\}$ and $d_{Q_3} = \max\{e_{Q_3} - \tilde{e}, |A\tilde{e}|\}$, which involve a constant A. The value $A = 0.05$ is recommended in Hidiroglou and Berthelot (1986). Ratios outside the interval

$$(\tilde{e} - Cd_{Q_1}, \tilde{e} + Cd_{Q_3}) \tag{9}$$

are flagged as outliers, where C will determine the width of the interval. Various values of C have been assessed in the literature; see Sigman (2002) and Thompson (2007). For our study, we use $C \in \{4, 10, 15\}$ since these provide a good representation of values found in

the literature. We note that an appropriate choice of U, A, and C is necessary before the procedure can be implemented.

The Hidiroglou-Berthelot method and the fence-based methods were both applied to microlevel ratio editing for the Annual Survey of Government Finances in Cornett et al. (2006). The authors found that the middle-fences rule and the Hidiroglou-Berthelot method provided better results for their application, which they explain is partly influenced by how the edit cells were formed. These methods (including some multivariate methods) were also investigated for macroediting using survey estimates from the U.S. Census Bureau's Annual Capital Expenditures Survey in Thompson (2007). That paper found that the Hidiroglou-Berthelot method performed the best, since it is designed to develop flexible limits when the ratios are "highly volatile." Thompson (2007) also underscores how it is difficult to develop a "one method fits all" approach to ratio editing, especially at the macrolevel. Thus it is important to emphasize that these methods, including the approach we present, are all possible tools for setting ratio edit tolerances and final determination should be done in coordination with a content-matter expert.

3. Statistical Tolerance Limits

By definition, a P/γ tolerance interval captures a specified proportion P (called the *content* of the tolerance interval) or more of a population with a given confidence level γ. A tolerance interval is computed using a random sample and the confidence level γ reflects the sampling variability. More formally, suppose a tolerance interval is to be computed for the distribution of a random variable X and let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ denote a random sample of size n. A P/γ two-sided tolerance interval, say $(L(\boldsymbol{X}), U(\boldsymbol{X}))$, computed using the random sample $\mathbf{X}$, satisfies

$$P_{\boldsymbol{X}}(P_X[L(\mathbf{X}) \leq X \leq U(\mathbf{X})|\mathbf{X}] \geq P) = \gamma. \tag{10}$$

The above condition states that with confidence level γ, the interval $(L(\mathbf{X}), U(\mathbf{X}))$ contains a proportion P or more of the distribution of X. As already noted, the confidence level γ reflects the sampling variability in the random sample $\mathbf{X}$. The quantities $L(\mathbf{X})$ and $U(\mathbf{X})$ are referred to as the tolerance limits. A one-sided tolerance interval, having only an upper or lower limit, can be similarly defined.

In this article, we use a two-sided tolerance interval for a normal distribution and a one-sided upper tolerance limit for a Weibull distribution. We shall now give expressions for the corresponding approximate tolerance limits. For a univariate normal distribution with unknown mean and unknown variance, let $\bar{X}$ and S^2 denote the sample mean and sample variance based on a sample of size n. Then a two-sided tolerance interval for the normal distribution is given by $\bar{X} \pm kS$, where the quantity k, referred to as a tolerance factor, has the approximate expression (see chap. 2 in Krishnamoorthy and Mathew 2009)

$$k = \left(\frac{(n-1)\chi^2_{1;P}(1/n)}{\chi^2_{n-1;1-\gamma}} \right)^{1/2}. \tag{11}$$

Here, $\chi^2_{1;P}(1/n)$ denotes the $100P$th percentile of a noncentral chi-square distribution with 1 degree of freedom (df) and noncentrality parameter $1/n$, while $\chi^2_{n-1;1-\gamma}$ denotes the $100(1-\gamma)$th percentile of a central chi-square distribution with $(n-1)$ df.

Now consider a random variable X following a Weibull distribution with scale parameter θ and shape parameter β, whose density is given by

$$f_X(x) = \frac{\beta}{\theta^\beta} x^{\beta-1} \exp\left[-\left(\frac{x}{\theta}\right)^\beta\right]. \tag{12}$$

Let $\hat{\theta}$ and $\hat{\beta}$ denote the maximum likelihood estimates of θ and β, respectively, based on a random sample of size n. An approximate P/γ upper tolerance limit for the Weibull distribution is given by

$$\exp\left\{\ln(\hat{\theta}) - \frac{t_{n-1;1-\gamma}\left(-\sqrt{n}\ln\{-\ln(1-P)\}\right)}{\hat{\beta}\sqrt{n-1}}\right\}, \tag{13}$$

where $t_{n-1;1-\gamma}\left(-\sqrt{n}\ln\{-\ln(1-P)\}\right)$ is the $100(1-\gamma)$th percentile of a non-central t distribution with (n − 1) df and non-centrality parameter $-\sqrt{n}\ln\{-\ln(1-P)\}$. The above approximation is due to Bain and Engelhardt (1981).

3.1 Statistical Tolerance Limits for Ratio Edits

If the data are roughly symmetric, an upper and lower tolerance bound may be needed to identify extremes in both tails of the data. However, ratio data are often right skewed. Thus, Thompson and Sigman (1999) suggest first omitting extreme observations of the untransformed data followed by a modified power transformation of the remaining data to obtain approximate symmetry.

There is some additional flexibility and insight gained by using statistical tolerance limits as an alternative to traditional ratio edit tolerance procedures. For example, we typically do not need to be concerned about transforming the data to near symmetry since approximate tolerance intervals have been developed for a wide range of distributions; see, for example, Krishnamoorthy and Mathew (2009). Also, the content and confidence levels of a tolerance interval allow us to reflect the uncertainty of what we are trying to capture with these intervals. Such uncertainty is not directly quantified by the traditional ratio edit tolerance procedures.

For the tolerance-limit approach, we first temporarily trim the data based on a user-specified trimming level. The assumption is that the remaining data behave similarly to the "true" uncontaminated distribution. The trimmed dataset is then used to calculate statistical tolerance limits, which can extend beyond the extremes of the trimmed data. Thus, some of the initially trimmed data may be retained as "good" data if they fall within the statistical tolerance limits, or the statistical tolerance limits may indicate that further data should be classified as ratio edit failures.

Another benefit to using statistical tolerance intervals is that the limits can never be negative for distributions with nonnegative support, regardless of the confidence and content levels specified. However, robust control limits and fence-based limits can yield negative lower bounds. While one can simply truncate the lower limits from these methods at zero, we do not have to specify this additional assumption when using statistical tolerance intervals.

4. Numerical Study

We now compare the performance of statistical tolerance limits with the traditional outlier procedures for determining ratio edits. All simulations in this section and calculations

for the example in the next section are performed using the R programming language (R Development Core Team 2013). Moreover, statistical tolerance limits are calculated using the R add-on package tolerance (Young 2010).

We compare the performance of the statistical tolerance limits with the ratio edit tolerances in two ways. First, we compute the average width of each procedure to comment on the relative conservatism of each procedure. Next, we compute the proportion of misclassified ratios with respect to each procedure's limits. We are interested in the proportion of false hits and misses, which are basically Type I and Type II error rates, respectively. Specifically, let X be a ratio. Then

$$\text{Type I Error Rate} = \Pr\{X \text{ flagged as "bad"} | X \text{ is "good"}\} \tag{14}$$

$$\text{Type II Error Rate} = \Pr\{X \text{ flagged as "good"} | X \text{ is "bad"}\} \tag{15}$$

Note that in the literature on outliers, the Type I and Type II errors defined above are rates of swamping and masking, respectively; we refer to Barnett and Lewis (1994) for further discussion on swamping and masking effects. We also note that some researchers may prefer to switch the definitions of Type I and Type II errors given above, unlike in a hypothesis-testing situation where Type I and Type II errors have universally accepted definitions. We chose the definitions given in (14) and (15) since they have already been used in the literature; cf. sec. 4.1 of Thompson and Sigman (1999).

In the case of a heavily skewed distribution, the region of outliers will typically be in the direction of the skewness. Therefore, instead of exploring simulated data where transformations could get the data close to symmetry, we will explore using one-sided trimming on the raw data in the direction of the skewness followed by a robust one-sided limit.

Our simulations assess the efficacy of one-sided tolerance limits and two-sided tolerance intervals for determining ratio edits. For the one-sided setting, we use a two-component mixture of Weibull distributions to simulate contamination in the upper tail of the data. For the two-sided setting, we use a three-component mixture of normals to simulate contamination in both tails of the data. It should be noted that mixture distributions (e.g., the contaminated normal model) have been used in the literature to assess the performance of editing procedures for survey data; see Ghosh-Dastidar and Schafer (2006). For each set of simulations, three scenarios were considered: well-separated components (i.e., a big gap between the "good" ratios and the "bad" ratios), moderate overlapping, and heavy overlapping.

Let Wei(θ, β) be the Weibull distribution with scale parameter θ and shape parameter β. Let N(μ,σ^2) be the normal distribution with mean μ and variance σ^2. The distributions we use for the one-sided contaminated simulations are:

- (Well Separated): 0.95*Wei(1,15) + 0.05*Wei(50,100)
- (Moderate Overlapping): 0.95*Wei(1,15) + 0.05*Wei(20,60)
- (Heavy Overlapping): 0.95*Wei(1,15) + 0.05*Wei(5,40)

The distributions we use for the two-sided contaminated simulations are:

- (Well Separated): 0.90*N(1000,$\sqrt{50}$) + 0.05*N(500,$\sqrt{50}$) + 0.05*N(1500,$\sqrt{50}$)

- (Moderate Overlapping): $0.90*\mathrm{N}(1000,\sqrt{50}) + 0.05*\mathrm{N}(750,\sqrt{50}) + 0.05*\mathrm{N}(1250,\sqrt{50})$
- (Heavy Overlapping): $0.90*\mathrm{N}(1000,\sqrt{50}) + 0.05*\mathrm{N}(900,\sqrt{50}) + 0.05*\mathrm{N}(1100,\sqrt{50})$

The following outlines the general simulation performed for our study:

1. Simulate n ratios, $X_1, \ldots, X_n$, from one of the contaminated models discussed above. Denote this sequence of ratios by $\mathbf{X}$.
2. Apply the traditional methods (i.e., the methods in Section 2) to $\mathbf{X}$ and calculate the ratio edit tolerances based on these approaches.
3. Use trimming at the $\alpha \in \{0.01, 0.02, \ldots, 0.10, 0.15\}$ levels on $\mathbf{X}$. Call these trimmed datasets $\mathbf{X}^{\alpha}$.
4. Using $\mathbf{X}^{\alpha}$, compute a normal statistical tolerance interval if contamination is assumed in both tails, or a one-sided upper Weibull statistical tolerance limit if contamination is assumed only in the right tail.
5. For each method and with respect to $\mathbf{X}$, calculate the proportion of good ratios falling outside of the tolerance limits (Type I error), and the proportion of bad ratios falling within the tolerance limits (Type II error).
6. Calculate the width of the statistical tolerance interval and the intervals determined by the traditional methods. For the one-sided setting, the one-sided upper tolerance limit will be taken as the width since an absolute lower limit of 0 is assumed for the data.
7. Repeat the above B times. For each method, average the Type I error rates, Type II error rates, and interval widths to get Monte Carlo estimates of each quantity.

For our simulations, we generate $n \in \{300, 1000\}$ ratios $B = 10{,}000$ times and compute P/γ tolerance intervals at the 90/90 and 95/95 levels. Recall from Section 3 that P is the content of the tolerance interval and γ is its confidence level. For the methods discussed in Section 2, we specify values for the constants (which we refer to as "Factors" in the summary tables) based on the references cited within.

Tables 1–3 give the simulation results for the three contamination structures considered for $n = 1{,}000$. The general results are similar for $n = 300$, which are reported in Tables 6–8 in the Appendix. We only report the results for a subset of the trimming levels used, but the trend in the average widths and errors as α changes is apparent. When the contamination structure is well separated or moderately overlaps with respect to the "good" data and a trimming level is selected close to the amount of contamination (5% for our simulations), then the statistical tolerance interval approach performs the best, namely meaning that the Type I error comes close to the nominal $(1 - \gamma)$ level. Note the results in bold in the tables, which pertain to the temporary trimming done at the true percentage of contamination. Regardless of the contamination structure, this approach does a good job of controlling the Type I errors as long as the level of trimming does not heavily exceed the content level P of the tolerance interval.

For the robust control limits, larger values of L yield smaller Type I errors, but larger Type II errors. Using $L \in \{2.0, 2.5, 3.0, 3.5\}$, we see there is generally a wide spread in the Type I and Type II errors. Again, we note that Thompson and Sigman (1999) use $L = 2$ for a more liberal rule and $L = 3$ for a more conservative rule regarding the number of cases

Table 1. Simulation results for the well-separated contamination structure (n = 1,000)

	One-sided			Two-sided		
Factors	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	29.6914	0.1383	0.0000	127.5013	0.2019	0.0000
2.5	34.3543	0.1011	0.0000	159.3766	0.1109	0.0000
3.0	39.0171	0.0741	0.0000	191.2519	0.0559	0.0000
3.5	43.6799	0.0543	0.0000	223.1272	0.0259	0.0000
k			Resistant fences			
1.5	51.5592	0.0323	0.0000	305.5729	0.0024	0.0000
2.0	60.9528	0.0174	0.0000	381.9661	0.0002	0.0000
3.0	79.7400	0.0050	0.0002	534.7526	0.0000	0.0000
*k**			Asymmetric fences			
3.0	59.8644	0.0188	0.0000	305.5729	0.0025	0.0000
4.0	72.0264	0.0085	0.0000	381.9661	0.0002	0.0000
6.0	96.3504	0.0017	0.3340	534.7526	0.0000	0.0000
(*U*, *C*)			Hidiroglou-Berthelot			
(0.3, 4)	50.4094	0.0361	0.0000	400.3468	0.0042	0.0000
(0.3, 10)	94.1750	0.0023	0.1147	781.5459	0.0000	0.0015
(0.3, 15)	155.5960	0.0003	0.9794	935.1183	0.0000	0.3112
(0.5, 4)	46.2717	0.0470	0.0000	401.3400	0.0041	0.0000
(0.5, 10)	80.9931	0.0057	0.0001	779.1278	0.0000	0.0006
(0.5, 15)	109.1108	0.0014	0.5276	925.9700	0.0000	0.2400
α			90/90 tolerance limits			
0.01	46.4975	0.0453	0.0000	496.0546	0.0000	0.0000
0.05	**36.3107**	**0.0895**	**0.0000**	**197.4260**	**0.0592**	**0.0000**
0.10	28.9472	0.1454	0.0000	131.8031	0.1873	0.0000
0.15	24.7898	0.1917	0.0000	108.7590	0.2763	0.0000
α			95/95 tolerance limits			
0.01	63.8249	0.0143	0.0000	596.0720	0.0000	0.0001
0.05	**48.2107**	**0.0407**	**0.0000**	**237.3212**	**0.0252**	**0.0000**
0.10	37.4035	0.0827	0.0000	158.5242	0.1126	0.0000
0.15	31.6277	0.1215	0.0000	130.8950	0.1901	0.0000

Table 2. Simulation results for the moderately overlapping contamination structure (n = 1,000)

	One-sided			Two-sided		
Factors	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	29.6914	0.1383	0.0000	127.5003	0.2019	0.0001
2.5	34.3542	0.1011	0.0000	159.3754	0.1109	0.0003
3.0	39.0170	0.0741	0.0002	191.2504	0.0559	0.0010
3.5	43.6799	0.0543	0.0024	223.1255	0.0259	0.0029
k			Resistant fences			
1.5	51.5592	0.0323	0.0679	305.5721	0.0024	0.0271
2.0	60.9528	0.0174	0.6918	381.9652	0.0002	0.1220
3.0	79.7400	0.0050	0.9797	534.7512	0.0000	0.6328
*k**			Asymmetric fences			
3.0	59.8644	0.0188	0.5919	305.5721	0.0025	0.0280
4.0	72.0264	0.0085	0.9773	381.9652	0.0002	0.1250
6.0	96.3504	0.0017	0.9797	534.7512	0.0000	0.6314
(*U*, *C*)			Hidiroglou-Berthelot			
(0.3, 4)	50.2838	0.0361	0.0408	303.1756	0.0042	0.0191
(0.3, 10)	102.4551	0.0023	0.9797	872.6594	0.0000	0.9098
(0.3, 15)	125.2692	0.0003	0.9866	1322.8198	0.0000	0.9798
(0.5, 4)	46.2483	0.0470	0.0078	303.7215	0.0041	0.0190
(0.5, 10)	81.8213	0.0057	0.9797	843.6258	0.0000	0.9296
(0.5, 15)	112.6471	0.0014	0.9797	1324.5677	0.0000	0.9798
α			90/90 tolerance limits			
0.01	40.2427	0.0685	0.0004	280.4024	0.0054	0.0146
0.05	**34.8773**	**0.0980**	**0.0000**	**171.6823**	**0.0870**	**0.0005**
0.10	28.9446	0.1454	0.0000	131.7994	0.1874	0.0001
0.15	24.7898	0.1917	0.0000	108.7582	0.2763	0.0001
α			95/95 tolerance limits			
0.01	53.6683	0.0280	0.1215	336.9388	0.0009	0.0535
0.05	**45.9078**	**0.0469**	**0.0064**	**206.3752**	**0.0398**	**0.0017**
0.10	37.3994	0.0827	0.0001	158.5197	0.1126	0.0003
0.15	31.6277	0.1215	0.0000	130.8940	0.1901	0.0001

Table 3. Simulation results for the heavily overlapping contamination structure (n = 1,000)

Factors	One-sided			Two-sided		
	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	28.9717	0.1450	0.1818	123.5955	0.2160	0.2226
2.5	33.4823	0.1071	0.3390	154.4944	0.1222	0.3247
3.0	37.9929	0.0793	0.5393	185.3932	0.0638	0.4421
3.5	42.5035	0.0588	0.7394	216.2921	0.0307	0.5644
k			Resistant fences			
1.5	50.9635	0.0336	0.9533	297.9491	0.0030	0.8351
2.0	60.2380	0.0182	0.9795	372.4364	0.0002	0.9535
3.0	78.7868	0.0053	0.9797	521.4109	0.0000	0.9798
*k**			Asymmetric fences			
3.0	58.9201	0.0199	0.9791	297.9491	0.0032	0.8341
4.0	70.8468	0.0091	0.9797	372.4364	0.0003	0.9521
6.0	94.7000	0.0019	0.9797	521.4109	0.0000	0.9798
(*U*, *C*)			Hidiroglou-Berthelot			
(0.3, 4)	49.5845	0.0377	0.9287	288.0104	0.0051	0.7931
(0.3, 10)	98.5630	0.0025	0.9797	1024.5677	0.0000	0.9798
(0.3, 15)	118.6151	0.0003	0.9962	1024.5677	0.0000	0.9836
(0.5, 4)	45.5290	0.0489	0.8369	288.2080	0.0050	0.7942
(0.5, 10)	80.1847	0.0062	0.9797	1024.5677	0.0000	0.9798
(0.5, 15)	106.9169	0.0015	0.9797	1024.5677	0.0000	0.9807
α			90/90 tolerance limits			
0.01	36.6272	0.0871	0.4750	184.5406	0.0649	0.4381
0.05	**31.9656**	**0.1188**	**0.2784**	**152.0293**	**0.1283**	**0.3155**
0.10	27.8326	0.1564	0.1511	125.9865	0.2075	0.2295
0.15	24.3998	0.1966	0.0817	105.4296	0.2912	0.1718
α			95/95 tolerance limits			
0.01	48.0975	0.0405	0.9146	221.7488	0.0266	0.5859
0.05	**41.4319**	**0.0631**	**0.6961**	**182.7508**	**0.0674**	**0.4318**
0.10	35.7166	0.0924	0.4345	151.5282	0.1293	0.3142
0.15	31.0461	0.1261	0.2475	126.8878	0.2040	0.2326

flagged for review. While the overall simulation results for $n = 300$ and $n = 1{,}000$ were similar, we note that the sample size does affect the errors for the robust control limits; that is, for larger n, the Type I error rates increase, while the Type II error rates decrease.

The fence-based methods are typically more conservative with respect to the statistical tolerance interval approach. As the contamination structure mixes more with the good data, we note that the Type II errors for the fence-based methods increase significantly with respect to the Type II errors for the tolerance intervals. We also note that the summaries are very similar for the two fence-based methods under the two-sided setting. This is expected given the symmetry of the generated data.

For most of the common values of the Hidiroglou-Berthelot method, we see that their performance is comparable to the statistical tolerance interval approach (at the 90/90 and 95/95 levels) under the well-separated case. The exceptions are when $(U, C) = (0.3, 10)$ and $(U, C) = (0.3, 15)$. Again, as the contamination structure mixes more with the good data, we note that the Type II errors increase significantly with respect to the Type II errors for the tolerance intervals.

Overall, the simulation results show that as more of the contaminated data mixes with the good data, masking becomes more prevalent. This results in intervals that do not (or cannot) exclude the contaminated data, which in turn increases the Type II errors for all procedures. When assessing the methods of Section 2, we simply used common levels found in the literature. Different results would obviously be obtained by adjusting the user-specified constants. But for a given set of data, the intuition may not always be apparent as to the trade-off in terms of the types of errors. However, the intuition with the values specified in the tolerance interval approach (i.e., α, P, and γ) are all clear. Informative choices of these levels will help control both types of errors, thus suggesting the utility of statistical tolerance intervals as a way to set ratio edit tolerances.

5. Annual Survey of Manufacturers

The Annual Survey of Manufacturers (ASM) collects data for the years between the Economic Census, which is conducted in the years ending in 2 and 7. The annual survey data are estimates derived from a statistically selected sample from all manufacturing establishments with one or more paid employees. The collection mode for this survey is through paper and internet reporting. Examples of statistics that the ASM reports for different manufacturing sectors include employment, payroll, operating expenses,value of shipments, and inventories.

In order to make the results of this example accessible and reproducible for the reader, our analysis uses the *Statistics for Industry Group and Industries* file for the years 2010 and 2011. The data can be accessed from the U.S. Census Bureau's website for the ASM found at `http://www.census.gov/manufacturing/asm/index.html`. The statistics are reported at various North American Industry Classification System (NAICS) levels. We use the lowest level reported, which is the six-digit NAICS industry grouping. We note that since this is officially published data, it has already gone through the U.S. Census Bureau's editing process. Our intent is to highlight the implementation of the statistical tolerance interval approach on this edited macrodata, which would typically be followed by a subject-matter expert's analysis of the flagged values.

We study six ratios for this example. Many of the variables comprising the ratios are reported in U.S. dollars, such as payroll, materials, and inventories. For all such quantities, the values are reported in $1,000 on the summary file. The ratios we study, as well as the abbreviations we use, are:

- **PR/NE**: annual payroll/number of employees;
- **MU/TS**: materials used/total value of shipments;
- **ME/MB**: materials and supplies at end of the year/materials and supplies at beginning of the year;
- **WH/WA**: all production worker's hours (in 1,000 hours)/production worker's average per year (i.e., the number of employees on payroll on certain days of the month specified by the ASM);
- **IE/IB**: total inventories at the end of the year/total inventories at the beginning of the year; and
- **WE/WB**: work-in-process inventories at the end of the year/work-in-process inventories at the beginning of the year.

The total number of industries for each dataset is 321. However, some ratios are not calculated since one or both of the values for an industry are withheld due to estimates not meeting publication or disclosure standards set by the U.S. Census Bureau.

We first determine whether a normal or Weibull distribution is most appropriate for each 5% trimmed ratio dataset. While we only explore these two distributions, there are no restrictions on which parametric distributions to investigate – especially if knowledge is available from a subject-matter expert. Regardless, we first use the Kolmogorov-Smirnov test to assess whether data from corresponding years follow the same distribution. Four of the ratios (PR/NE, MU/TS, ME/MB, and IE/IB) yield p-values well over 0.15, while the other two ratios (WH/WA and WE/WB) have p-values below 0.05.

For the four ratios that have statistically similar distributions between the two years, we temporarily pool each pair of ratio datasets. We use the Shapiro-Wilk test for normality and the chi-square goodness-of-fit for testing the Weibull assumption. We then select the distribution of which corresponding test had the higher p-value. While these are two different tests, this is merely a simple approach to decide upon a distribution.

For the two ratios that are significantly different, we proceed similarly with testing the normality or the Weibull assumption. However, we keep each year's data separate and run the tests on these datasets. We then choose the distribution of which the test yielded the higher p-value between the two datasets for a given ratio.

After determining to proceed with the Weibull or normal assumption, we then compute one-sided tolerance limits or two-sided tolerance intervals, respectively. We consider the 90/90 and 95/95 levels with an initial trimming of 5%. We also perform a relative comparison between the 2010 and 2011 ratios. Specifically, we compare the proportions of how an industry is classified (i.e., as being "good" or an "outlier") from 2010 and 2011. These quantities give us an indication of how stable the classifications are from 2010 to 2011 with respect to the calculated limits.

For the PR/NE ratios, we also calculate the other limits discussed in this article. We found that the Weibull distribution is appropriate for both the 2010 and 2011 data. Thus, we calculate 90/90 and 95/95 one-sided upper Weibull tolerance limits. The results are

Table 4. Comparison of the one-sided upper limits for the PR/NE data

Method	Upper limit (2010)	Upper limit (2011)	Outlier to good	Good to outlier	Good to good
L			Robust control limits		
2.0	68.9846	71.0740	0.0031	0.0094	0.8840
2.5	73.7211	76.0442	0.0125	0.0031	0.9154
3.0	78.4577	81.0143	0.0031	0.0031	0.9498
3.5	83.1942	85.9844	0.0063	0.0031	0.9687
k			Resistant fences		
1.5	84.5603	86.1832	0.0000	0.0063	0.9749
2.0	92.9164	94.7511	0.0000	0.0000	0.9875
3.0	109.6287	111.8871	0.0000	0.0000	1.0000
*k**			Asymmetric fences		
3.0	88.3639	88.8199	0.0031	0.0031	0.9812
4.0	97.9879	98.2668	0.0000	0.0000	0.9875
6.0	117.2360	117.1606	0.0000	0.0000	1.0000
(*U*, *C*)			Hidiroglou-Berthelot		
(0.3, 4)	87.9928	87.2365	0.0000	0.0063	0.9781
(0.3, 10)	105.6065	107.3582	0.0000	0.0000	1.0000
(0.3, 15)	105.6065	107.3582	0.0000	0.0000	1.0000
(0.5, 4)	83.0819	84.1461	0.0063	0.0157	0.9561
(0.5, 10)	105.6065	107.3582	0.0000	0.0000	1.0000
(0.5, 15)	105.6065	107.3582	0.0000	0.0000	1.0000
P/γ			Tolerance limits		
90/90	65.8670	67.7087	0.0157	0.0094	0.8464
95/95	70.1582	72.1881	0.0063	0.0094	0.8934

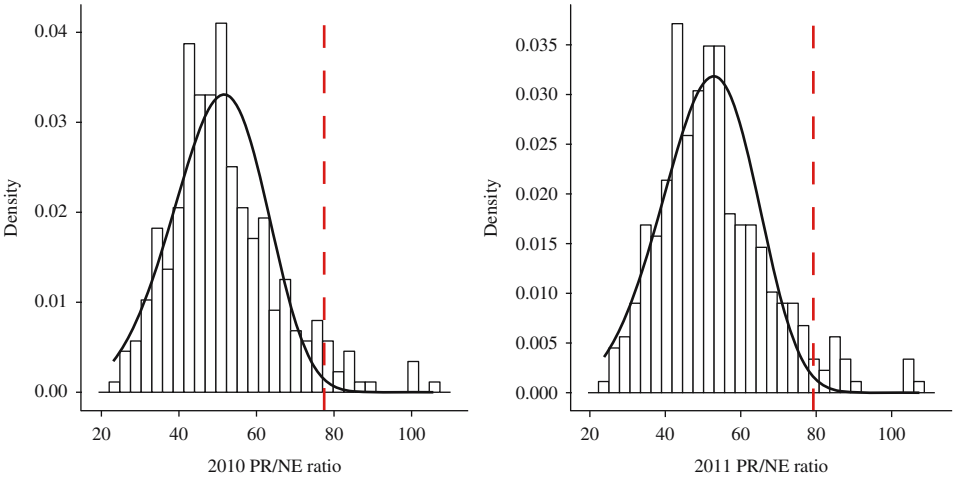

Fig. 1. Histograms of the PR/NE ratios for (a) 2010 and (b) 2011. The dashed line represents the 5% trimming threshold and the solid line is the Weibull density curve fit to the trimmed data

reported in Table 4. We see that the resistant fences provide fairly conservative limits. As such, the proportion of points classified as "good" to "good" is close to or at 1 and this conservatism is likely not desirable. As the histograms in Figure 1 show, there are clearly a few ratios above the value of 90 that may be candidates for editing. The robust control limits and the tolerance interval procedures would flag these values for possible editing, whereas the other approaches produce fairly conservative limits. Given the ability to better control Type I and Type II errors with the statistical tolerance intervals, their use here gives this approach a significant edge over the other procedures.

Scatterplots of the payroll versus the number of employees for each year are given in Figure 2. As can be seen, each year shows a strong correlation (which is approximately $+0.93$ for each year). Values flagged using the 90/90 and 95/95 tolerance limits are color coded accordingly. One thing to note is that as the correlation strengthens, the resulting tolerance limits will be "tighter" around the data.

Results for the other five ratios are similar to those reported for the PR/NE ratios. Hence, we only focus on the tolerance interval results. Table 5 gives the one-sided tolerance limit or two-sided tolerance interval results depending on the distributional assumption made. For the 90/90 limits, approximately 70% to 85% of the data stay within the limits across years, while for the 95/95 limits, these same percentages range from approximately 80% to 90%. These percentages give an indication of those industries that have essentially remained stable between 2010 and 2011. If one wants to develop certain summary statistics between the two years, then those industries that fell outside of the limits in one or both years could be candidates for editing. Moreover, they could be indicative of changes that occurred within that particular industry.

6. Discussion

The criterion used in developing a statistical tolerance interval indicates that it is a natural choice for computing bounds that can be used to perform ratio edits; that is, in order to flag ratios that are inconsistent or problematic. In our work, we have demonstrated this in the

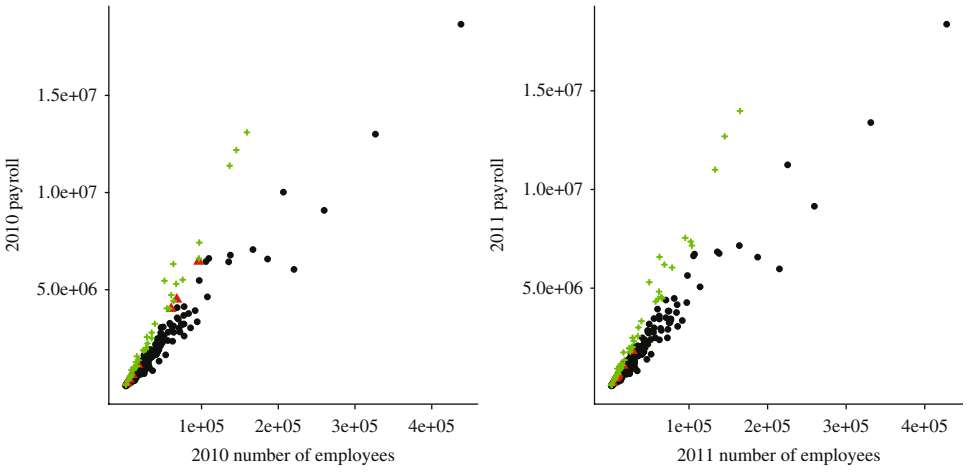

Fig. 2. Scatterplots of the payroll (in $1,000) versus the number of employees for (a) 2010 and (b) 2011. The triangles are values greater than the 95/95 upper tolerance limit, while the plusses and triangles are values greater than the 90/90 upper tolerance limit

case of the normal distribution (where the problematic ratios can appear in either tail of the distribution) and in the case of the Weibull distribution (where the problematic ratios appear only in the right tail). A comparison with other ratio edit procedures shows that the statistical tolerance-interval approach has a significant edge over the existing procedures in terms of controlling Type I and Type II errors. The approach also depends on an initial level of trimming. As noted in Section 1, there is no clear guidance on choosing a percentage of trimming to perform, so one should seek input from a subject-matter expert. Our approach can certainly be adopted for other distributions; see Krishnamoorthy and Mathew (2009) for details on the development of tolerance intervals for a variety of distributions.

We also acknowledge that the ratio editing process is often complex and includes numerous rules that are typically dependent on the type of survey. Moreover, ratio editing at the microlevel and macrolevel often use different approaches, with the latter setting not as well studied in the literature. We illustrated the statistical tolerance interval approach on ASM data at the macrolevel, but the approach is applicable to the microlevel setting. We are not suggesting a panacea for setting ratio edit tolerances in all survey settings; however, we are suggesting that statistical tolerance intervals can be useful in informing ratio editing processes.

We note that both of the variables used in the computation of a ratio can have values that are outliers, and yet the ratio will not be flagged as an outlier. This can obviously happen when values of both variables are too small or too large, so that the outlyingness gets cancelled when we take the ratio. A simple example is if a small business reports 400 trillion dollars in payroll for ten million employees, then the PR/NE ratio would be consistent with those displayed in Figure 1. In view of this, it is essential to have outlier detection methods that are applicable to bivariate data, or to multivariate data when data are available on several variables. A Mahalanobis distance based outlier detection method (cf. Franklin et al. (2000) and Thompson (2007)) may not adequately flag the outliers, since the outlyingness of a single variable (or a few variables) may be cancelled out by the magnitudes of the other variables. We believe a rectangular tolerance region that provides simultaneous tolerance intervals on each variable is required. Such a tolerance region is currently under investigation.

Table 5. Comparison of the statistical tolerance intervals for the other five ratio datasets. The second column gives the distributional assumption, which determined whether we calculated a one-sided tolerance limit or two-sided tolerance interval

Ratio	Dist.	P/γ	2010 limits	2011 limits	Outlier to good	Good to outlier	Good to good
MU/TS	Weibull	90/90	0.5914	0.6053	0.0221	0.0063	0.8360
		95/95	0.6385	0.6546	0.0126	0.0189	0.8991
ME/MB	Normal	90/90	(0.9339, 1.2204)	(0.9555, 1.2504)	0.1169	0.1039	0.7143
		95/95	(0.9037, 1.2507)	(0.9243, 1.2815)	0.0714	0.0649	0.8149
WH/WA	Weibull	90/90	2.0850	2.0952	0.0313	0.0251	0.8464
		95/95	2.1065	2.1182	0.0157	0.0251	0.8903
IE/IB	Normal	90/90	(0.9461, 1.2143)	(0.9712, 1.2186)	0.1266	0.1266	0.6899
		95/95	(0.9178, 1.2425)	(0.9452, 1.2447)	0.0791	0.1076	0.7911
WE/WB	Normal	90/90	(0.8513, 1.3252)	(0.9068, 1.2868)	0.1303	0.1303	0.6775
		95/95	(0.8013, 1.3752)	(0.8667, 1.3269)	0.0912	0.0879	0.7818

Appendix: Additional Simulation Results

Table 6. Simulation results for the well-separated contamination structure ($n = 300$)

	One-sided			Two-sided		
Factors	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	35.1594	0.0950	0.0000	127.8109	0.1996	0.0000
2.5	41.1869	0.0638	0.0000	159.7636	0.1095	0.0000
3.0	47.2145	0.0429	0.0000	191.7163	0.0554	0.0000
3.5	53.2420	0.0289	0.0000	223.6690	0.0258	0.0000
k			Resistant fences			
1.5	51.3763	0.0330	0.0000	304.5684	0.0028	0.0000
2.0	60.7281	0.0179	0.0000	380.7105	0.0002	0.0000
3.0	79.4316	0.0053	0.0089	532.9946	0.0000	0.0000
*k**			Asymmetric fences			
3.0	59.5974	0.0197	0.0000	304.5684	0.0033	0.0000
4.0	71.6895	0.0091	0.0015	380.7105	0.0003	0.0000
6.0	95.8738	0.0020	0.3582	532.9946	0.0000	0.0000
(*U*, *C*)			Hidiroglou-Berthelot			
(0.3, 4)	51.4053	0.0370	0.0000	595.6948	0.0051	0.0000
(0.3, 10)	101.8578	0.0026	0.2108	833.5057	0.0000	0.0039
(0.3, 15)	156.3003	0.0004	0.8997	1000.5404	0.0000	0.3038
(0.5, 4)	46.9528	0.0480	0.0000	596.8402	0.0050	0.0000
(0.5, 10)	85.1081	0.0062	0.0083	831.7058	0.0000	0.0018
(0.5, 15)	118.1210	0.0016	0.4681	954.3806	0.0000	0.2579
α			90/90 tolerance limits			
0.01	48.5763	0.0399	0.0000	507.6096	0.0000	0.0000
0.05	**38.0012**	**0.0808**	**0.0000**	**223.7015**	**0.0459**	**0.0000**
0.10	30.1621	0.1342	0.0000	136.2889	0.1721	0.0000
0.15	25.7931	0.1790	0.0000	112.6394	0.2582	0.0000

Table 7. Simulation results for the moderately overlapping contamination structure (n = 300)

Factors	One-sided			Two-sided		
	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	35.1591	0.0950	0.0001	127.8100	0.1996	0.0001
2.5	41.1866	0.0638	0.0021	159.7626	0.1095	0.0003
3.0	47.2141	0.0429	0.0284	191.7151	0.0554	0.0011
3.5	53.2415	0.0289	0.1795	223.6676	0.0258	0.0032
k			Resistant fences			
1.5	51.3763	0.0330	0.1136	304.5681	0.0028	0.0285
2.0	60.7281	0.0179	0.6064	380.7101	0.0002	0.1264
3.0	79.4316	0.0053	0.9270	532.9941	0.0000	0.6217
*k**			Asymmetric fences			
3.0	59.5974	0.0197	0.5271	304.5681	0.0033	0.0313
4.0	71.6895	0.0091	0.8880	380.7101	0.0003	0.1349
6.0	95.8738	0.0020	0.9286	532.9941	0.0000	0.6171
(*U*, *C*)			Hidiroglou-Berthelot			
(0.3, 4)	50.7645	0.0370	0.0822	329.2378	0.0051	0.0211
(0.3, 10)	111.4427	0.0026	0.9280	1020.4076	0.0000	0.8680
(0.3, 15)	123.4679	0.0004	0.9561	1270.2145	0.0000	0.9313
(0.5, 4)	46.7444	0.0480	0.0197	329.7336	0.0050	0.0210
(0.5, 10)	92.8964	0.0062	0.9232	1056.5267	0.0000	0.8836
(0.5, 15)	117.3996	0.0016	0.9287	1271.0021	0.0000	0.9311
α			90/90 tolerance limits			
0.01	42.0401	0.0607	0.0015	287.8207	0.0050	0.0190
0.05	**36.2829**	**0.0891**	**0.0001**	**180.5547**	**0.0740**	**0.0007**
0.10	30.1462	0.1343	0.0000	136.2762	0.1721	0.0001
0.15	25.7931	0.1790	0.0000	112.6387	0.2582	0.0001
α			95/95 tolerance limits			
0.01	57.1274	0.0221	0.3816	348.3759	0.0008	0.0717
0.05	**48.6076**	**0.0393**	**0.0386**	**218.6988**	**0.0310**	**0.0030**
0.10	39.6391	0.0711	0.0010	165.2392	0.0975	0.0004
0.15	33.4736	0.1069	0.0000	136.7524	0.1700	0.0002

Table 8. Simulation results for the heavily overlapping contamination structure ($n = 300$)

	One-sided			Two-sided		
Factors	Average width	Type I error	Type II error	Average width	Type I error	Type II error
L			Robust control limits			
2.0	33.7435	0.1043	0.3568	123.7923	0.2140	0.2261
2.5	39.4462	0.0714	0.6078	154.7404	0.1210	0.3294
3.0	45.1488	0.0490	0.8155	185.6885	0.0634	0.4452
3.5	50.8515	0.0336	0.9080	216.6366	0.0308	0.5674
k			Resistant fences			
1.5	50.7531	0.0343	0.8999	296.9006	0.0035	0.8242
2.0	59.9802	0.0187	0.9272	371.1257	0.0003	0.9164
3.0	78.4344	0.0056	0.9280	519.5760	0.0000	0.9311
k^*			Asymmetric fences			
3.0	58.6094	0.0208	0.9251	296.9006	0.0040	0.8205
4.0	70.4553	0.0098	0.9279	371.1257	0.0004	0.9129
6.0	94.1470	0.0022	0.9323	519.5760	0.0000	0.9310
(U, C)			Hidiroglou-Berthelot			
(0.3, 4)	50.3828	0.0387	0.8765	378.4962	0.0061	0.7847
(0.3, 10)	103.0961	0.0028	0.9291	971.1992	0.0000	0.9311
(0.3, 15)	112.2477	0.0004	0.9825	971.2298	0.0000	0.9518
(0.5, 4)	45.9667	0.0500	0.8010	377.1936	0.0060	0.7854
(0.5, 10)	88.3975	0.0066	0.9280	971.2298	0.0000	0.9311
(0.5, 15)	107.4762	0.0017	0.9344	971.2298	0.0000	0.9457
α			90/90 tolerance limits			
0.01	38.3546	0.0773	0.5560	189.7963	0.0575	0.4602
0.05	**33.2305**	**0.1088**	**0.3305**	**156.6118**	**0.1165**	**0.3346**
0.10	28.9141	0.1452	0.1825	130.0538	0.1921	0.2448
0.15	25.3551	0.1842	0.0987	109.1294	0.2730	0.1831
α			95/95 tolerance limits			
0.01	51.3207	0.0324	0.9143	229.7280	0.0213	0.6194
0.05	**43.8078**	**0.0535**	**0.7789**	**189.6976**	**0.0571**	**0.4623**
0.10	37.7291	0.0801	0.5315	157.6943	0.1135	0.3400
0.15	32.8047	0.1116	0.3196	132.4919	0.1838	0.2532

7. References

Aggarwal, C.C. 2013. *Outlier Analysis*. New York: Springer.

Bain, L.J. and M. Engelhardt. 1981. "Simple Approximate Distributional Results for Confidence and Tolerance Limits for the Weibull Distribution Based on Maximum Likelihood Estimators." *Technometrics* 23: 15–20. DOI: http://dx.doi.org/10.1080/00401706.1981.10486231

Barnett, V. and T. Lewis. 1994. *Outliers in Statistical Data*, 3rd ed. Wiley Series in Probability and Mathematical Statistics. Chichester: John Wiley & Sons.

Belcher, R. 2003. "Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys." In Proceedings of the Survey Methods Section: Statistical Society of Canada Annual Meeting, June, 2003. 25–30 Halifax, Nova Scotia, Canada. Available at: http://www.ssc.ca/survey/documents/SSC2003_R_Belcher.pdf

Chawla, S. and A. Gionis. 2013. "*k*-means-: A Unified Approach to Clustering and Outlier Detection." In Proceedings of the 2013 SIAM International Conference on Data Mining: Society for Industrial and Applied Mathematics, May, 2013. 187–197 Austin, Texas, USA. Available at http://epubs.siam.org/doi/pdf/10.1137/1.9781611972832.21

Cornett, E., J.F. McLaughlin, and C.R. Hogue. 2006. "A Comparison of Two Ratio Edit Methods for the Annual Survey of Government Finances." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, 2006. 2878–2883 Seattle, WA, USA. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2006/Files/JSM2006-000199.pdf

Franklin, S., S. Thomas, and M. Brodeur. 2000. "Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Modified Stahel-Donoho Estimators." In Proceedings of the Second International Conference on Establishment Surveys (ICES-II), Survey Methods for Businesses, Farms, and Institutions, June, 2000. 697–706 Buffalo, NY, USA. Available at: http://www.amstat.org/meetings/ices/2000/proceedings/S33.pdf

Ghosh-Dastidar, B. and J.L. Schafer. 2006. "Outlier Detection and Editing Procedures for Continuous Multivariate Data." *Journal of Official Statistics* 22: 487–506.

Hidiroglou, M.A. and J.-M. Berthelot. 1986. "Statistical Editing and Imputation for Periodic Business Surveys." *Survey Methodology* 12: 73–83.

Hido, S., Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. 2011. "Statistical Outlier Detection Using Direct Density Ratio Estimation." *Knowledge and Information Systems* 26: 309–336. DOI: http://dx.doi.org/10.1007/s10115-010-0283-2

Hoaglin, D.C., B. Iglewicz, and J.W. Tukey. 1986. "Performance of Some Resistant Rules for Outlier Labeling." *Journal of the American Statistical Association* 81: 991–999. DOI: http//dx.doi.org/10.1080/01621459.1986.10478363

Iglewicz, B. and D.C. Hoaglin. 1993. *How to Detect and Handle Outliers*, vol. 16. Milwaukee, WI: American Society for Quality Control.

Kokic, P.N. and P.A. Bell. 1994. "Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator." *Journal of Official Statistics* 10: 419–435.

Krishnamoorthy, K. and T. Mathew. 2009. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Hoboken, NJ: Wiley.

Latouche, M. and J.-M. Berthelot. 1992. "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys." *Journal of Official Statistics* 8: 389–400.

Montgomery, D.C. 2013. *Introduction to Statistical Quality Control*, 7th ed. Hoboken, NJ: Wiley.

R Development Core Team. 2014. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ ISBN 3-900051-07-0 (accessed February 13, 2015)

Rais, S. 2008. "Outlier Detection for the Consumer Price Index." In Proceedings of the Survey Methods Section: Statistical Society of Canada Annual Meeting, May, 2008. 1–10 Ottawa, Ontario, Canada. Available at: http://www.ssc.ca/survey/documents/SSC2068_5_Rais.pdf.

Rivest, L.-P. and M. Hidiroglou. 2004. "Outlier Treatment for Disaggregated Estimates." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, 2004. 4248–4256 Toronto, Ontario, Canada. Available at: http://www.amstat.stat.org/sections/SRMS/Proceedings/y2004/files/Jsm2004-000149.pdf

Rousseeuw, P. and A.M. Leroy. 2003. *Robust Regression and Outlier Detection*. Hoboken, NJ: Wiley Series in Probability and Mathematical Statistics.

Shewhart, W.A. 1939. *Statistical Method from the Viewpoint of Quality Control*. Washington, DC: Dover.

Sigman, R.S. 2002. "Statistical Methods Used to Detect Cell-Level and Respondent-Level Outliers in the 2002 Economic Census of the Services Sector." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 2002. 3566–3573 Minneapolis, MN, USA. Available at: https://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSM2005-000465.pdf

Tambay, J.-L. 1988. "An Integrated Approach for the Treatment of Outliers in Sub-Annual Economic Surveys." In Proceedings of the Section on Survey Research Methods: American Statistical Association, 229–234. Available at: http://www.amstat.org/sections/SRMS/Proceedings/papers/1988-040.pdf

Thompson, K.J. 1999. "Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods." In Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference, November 1999. 1–10. Arlington, VA, USA. Available at: https://fcsm.sites.usa.gov/files/2014/05/VII-B_Thompson_FCSM1999.pdf.

Thompson, K.J. 2007. "Investigation of Macro Editing Techniques for Outlier Detection in Survey Data." In Proceedings of the Third International Conference on Establishment Surveys (ICES-III), Survey Methods for Businesses, Farms, and Institutions, June 2007. 1186–1193. Montreal, Quebec, Canada. Available at http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000071.pdf.

Thompson, K.J. and S.A. Adeshiyan. 2003. "Data Quality Effects of Alternative Edit Parameters." *Journal of Data Science* 1: 1–25.

Thompson, K.J. and R.S. Sigman. 1999. "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data." *Journal of Official Statistics* 15: 517–535.

Tukey, J.W. and D.H. McLaughlin. 1963. "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1." *Sankhyā: The Indian Journal of Statistics (Series A)* 25: 331–352.

Young, D.S. 2010. "Tolerance: An R package for Estimating Tolerance Intervals." *Journal of Statistical Software* 36: 1–39. Available at: http://www.jstatsoft.org/v36/i05/ (accessed February 13, 2015)

Yuen, K.-V. and H.-Q. Mu. 2012. "A Novel Probabilistic Method for Robust Parametric Identification and Outlier Detection." *Probabilistic Engineering Mechanics* 30: 48–59. DOI: http://dx.doi.org/10.1016/j.probengmech.2012.06.002

Received August 2013
Revised April 2014
Accepted June 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 101–119, http://dx.doi.org/10.1515/JOS-2015-0005

On Estimating Quantiles Using Auxiliary Information

Yves G. Berger[1] *and Juan F. Muñoz*[2]

We propose a transformation-based approach for estimating quantiles using auxiliary information. The proposed estimators can be easily implemented using a regression estimator. We show that the proposed estimators are consistent and asymptotically unbiased. The main advantage of the proposed estimators is their simplicity. Despite the fact the proposed estimators are not necessarily more efficient than their competitors, they offer a good compromise between accuracy and simplicity. They can be used under single and multistage sampling designs with unequal selection probabilities. A simulation study supports our finding and shows that the proposed estimators are robust and of an acceptable accuracy compared to alternative estimators, which can be more computationally intensive.

Key words: Distribution function; inclusion probabilities; regression estimator; sample survey.

1. Introduction

Estimation of quantiles is of considerable interest when measuring income distribution and poverty lines (e.g. Osier 2009; Verma and Betti 2011; Eurostat 2003; Berger and Skinner 2003). For instance, the median is regarded as a more appropriate measure of location than the mean when variables of interest, such as income, expenditure, and so on, have highly skewed distributions, because the median is less sensitive to outliers than the mean. For this reason, the median is also used by most household wealth surveys, such as the Household Finance and Consumption Survey (HFCS) carried out by the European Central Bank among the Eurozone countries. In addition, quantile estimation has many practical applications, for example, when measuring poverty (e.g. Osier 2009; Eurostat 2012; Eurostat 2003).

In sample surveys, auxiliary information is often used at the estimation stage to improve the estimation of target parameters. The use of auxiliary information has been studied extensively for estimation of means and totals. However, it has no obvious extensions to the estimation of quantiles. In this article, we propose a transformation-based approach for estimating quantiles, which takes into account of the auxiliary information.

We consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ containing N units. Let $y_1, \ldots, y_N$ denote the values of a variable of interest, y, and $x_1, \ldots, x_N$ denote the values of an auxiliary variable, x. Our proposed approach can be easily extended to several auxiliary variables. A sample s of size n is selected randomly from U according to

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK. Email: y.g.berger@soton.ac.uk
[2] Department of Quantitative Methods in Economics and Business, University of Granada, Granada, 18071, Spain. Email: jfmunoz@ugr.es

a sampling design. We consider a design-based approach where the y_i and x_i are fixed (nonrandom) quantities and the sampling distribution is specified by the sampling design. The aim is to estimate the population quantile

$$Y_\alpha = F^{-1}(\alpha), \tag{1}$$

where $F^{-1}(\cdot)$ is the inverse of the population distribution function

$$F(t) = \frac{1}{N}\sum_{i\in U}\delta(y_i \le t)$$

and $0 < \alpha < 1$. The function $\delta(\cdot)$ takes the value 1 if its argument is true and 0 otherwise. Throughout this article, we define the inverse of any function $G(\cdot)$ by $G^{-1}(\alpha) = inf\{t : G(t) \ge \alpha\}$.

A customary estimator for Y_α is obtained by substituting $F(t)$ by its estimator into (1). For example, the 'Hájek type' estimator of Y_α is defined by

$$\widehat{Y}_{\pi,\alpha} = \widehat{F}_\pi^{-1}(\alpha), \tag{2}$$

where $\widehat{F}_\pi(t)$ is the Hájek (1971) estimator defined by

$$\widehat{F}_\pi(t) = \frac{1}{\widehat{N}}\sum_{i\in s}\frac{1}{\pi_i}\delta(y_i \le t) \tag{3}$$

with $\widehat{N} = \sum_{i\in s}\pi_i^{-1}$, where π_i denotes the first-order inclusion probability of unit i. A wide range of estimators exists for the distribution function $F(\cdot)$, some of which use auxiliary information (see Section 2).

The proposed approach consists in inverting the distribution function at $\widehat{\alpha}_{reg}$ rather than at α. The quantity $\widehat{\alpha}_{reg}$, defined in (19), takes the auxiliary information into account. The proposed estimators can be justified by using a transformation of the variable of interest. The proposed estimators depend on the first-order inclusion probabilities. The proposed estimators can be calculated even if we only know the auxiliary variables for the sampled units, as long as the population quantile of the auxiliary variable is known.

In Section 2, we define estimators of the distribution function that can be found in the literature, and which can be used to estimate a quantile. In Section 3, we introduce the proposed estimators for a quantile. In Section 4, we give regularity conditions under which the proposed estimators are consistent. In Section 5, we compare the proposed estimators with alternative estimators via simulation. We also investigate the empirical properties of a bootstrap variance estimator. This article concludes with some discussions in Section 6.

2. Estimators of Quantiles

An exhaustive review of estimators of the distribution function and quantiles can be found in Dorfman (2009).

By substituting the design weights in (3) with calibration weights, we obtain the following naïve estimator

$$\widehat{F}_w(t) = \frac{1}{\widehat{N}_w} \sum_{i \in s} w_i \delta(y_i \leq t), \tag{4}$$

where $\widehat{N}_w = \sum_{i \in s} w_i$. The w_i denote the regression weights calibrated with respect to the population total of the auxiliary variable. The estimator of Y_α based on these calibration weights is given by $\widehat{Y}_{w;\alpha} = \widehat{F}_w^{-1}(\alpha)$.

The model-based estimator of the distribution function suggested by Chambers and Dunstan (1986) is based on the following heteroscedastic regression model

$$y_i = \beta x_i + \nu(x_i) u_i, \tag{5}$$

where β is an unknown parameter, $\nu(x_i)$ is a known function of x and the u_i are independent and identically distributed random variables with zero mean. The distribution function estimator proposed by Chambers and Dunstan (1986) is

$$\widehat{F}_{cd}(t) = \left[\sum_{i \in s} \delta(y_i \leq t) + \frac{1}{n_j} \sum_{j \in U-s} \sum_{i \in s} \delta\left(u_{ni} \leq \frac{t - b_n x_j}{\nu(x_j)} \right) \right], \tag{6}$$

with

$$b_n = \left[\sum_{i \in s} \frac{x_i^2}{\nu^2(x_i)} \right]^{-1} \sum_{i \in s} \frac{y_i x_i}{\nu^2(x_i)}; \quad u_{ni} = \frac{y_i - b_n x_i}{\nu(x_i)}.$$

The Chambers and Dunstan (1986) estimator of Y_α is given by $\widehat{Y}_{cd;\alpha} = \widehat{F}_{cd}^{-1}(\alpha)$.

Rao et al. (1990) proposed the following estimator

$$\widehat{F}^{\bullet}_{rkm}(t) = \frac{1}{N} \left\{ \sum_{i \in s} \pi_i^{-1} \delta(y_i \leq t) + \left(\sum_{i \in U} \widehat{G}_i(t) - \sum_{i \in s} \pi_i^{-1} \widehat{G}_{ic}(t) \right) \right\}$$

with

$$\widehat{G}_i(t) = \frac{1}{\widehat{N}} \sum_{j \in s} \frac{1}{\pi_j} \delta\left(\widehat{u}_j \leq \frac{t - \widehat{R} x_i}{x_i^{1/2}} \right),$$

$$\widehat{G}_{ic}(t) = \left(\sum_{j \in s} \frac{\pi_i}{\pi_{ij}} \right)^{-1} \left[\sum_{j \in s} \frac{\pi_i}{\pi_{ij}} \delta\left(\widehat{u}_j \leq \frac{t - \widehat{R} x_i}{x_i^{1/2}} \right) \right],$$

$$\widehat{u}_j = \frac{y_j - \widehat{R} x_j}{x_j^{1/2}}, \widehat{R} = \left[\sum_{i \in s} \frac{x_i}{\pi_i} \right]^{-1} \sum_{i \in s} \frac{y_i}{\pi_i},$$

where π_{ij} denotes the joint inclusion probability for the units i and j. Since the estimator $\widehat{F}^{\bullet}_{rkm}(t)$ is not always a monotone nondecreasing function, Rao et al. (1990) proposed to use the following estimator

$$\widehat{F}_{rkm}(t) = \max\left\{\widetilde{F}_{rkm}(y_{(i)}) : y_{(i)} \leq t\right\}, \tag{7}$$

where the $y_{(i)}$'s are the order statistics of the sample $\{y_i, i \in s\}$ and $\widetilde{F}_{rkm}(y_{(i)})$ is defined by the following recursive formula

$$\widetilde{F}_{rkm}(y_{(i)}) = \max\left\{\widetilde{F}_{rkm}(y_{(i-1)}), \widehat{F}^{\bullet}_{rkm}(y_{(i)})\right\},$$

with $\widetilde{F}_{rkm}(y_{(1)}) = \widehat{F}^{\bullet}_{rkm}(y_{(1)})$. The Rao et al. (1990) estimator of Y_α is given by $\widehat{Y}_{rkm;\alpha} = \widehat{F}^{-1}_{rkm}(\alpha)$.

Silva and Skinner (1995) proposed the following estimator based on poststratification

$$\widehat{F}_{ps}(t) = \frac{1}{N}\sum_{g=1}^{G}\frac{N_g}{\widehat{N}_g}\sum_{i\in s}\frac{1}{\pi_i}\delta(y_i \leq t)\delta(i \in U_g), \tag{8}$$

where $U_1, \ldots, U_G$ are G poststrata partitioning the population, N_g is the size of U_g and $\widehat{N}_g = \sum_{i\in s_g}\pi_i^{-1}$, with $g = 1, \ldots, G$. The estimator of Y_α is given by $\widehat{Y}_{ps;\alpha} = \widehat{F}^{-1}_{ps}(\alpha)$.

When the population quantile X_α of an auxiliary variable is known, Rao et al. (1990) proposed the following ratio estimator of Y_α

$$\widehat{Y}_{r;\alpha} = \frac{\widehat{Y}_{\pi;\alpha}}{\widehat{X}_{\pi;\alpha}}X_\alpha, \tag{9}$$

where $\widehat{Y}_{\pi;\alpha}$ and $\widehat{X}_{\pi;\alpha}$ are respectively the Hájek estimators of Y_α and X_α (see (2)). Rao et al. (1990) also proposed a difference estimator and showed that $\widehat{Y}_{r;\alpha}$ has a smaller mean square error than the difference estimator.

Harms and Duchesne (2006) proposed an estimator of the distribution function based on a calibration constraint specified by the quantile of an auxiliary variable. This estimator is denoted by $\widehat{Y}_{cal;\alpha}$.

Note that the estimators $\widehat{Y}_{cd;\alpha}$, $\widehat{Y}_{rkm;\alpha}$ and $\widehat{Y}_{ps;\alpha}$ assume that the auxiliary variable is known for all the units of the population, whereas estimators $\widehat{Y}_{r;\alpha}$ and $\widehat{Y}_{cal;\alpha}$ only require the knowledge of X_α.

3. Proposed Estimators for a Quantile

The proposed estimators are based upon the following idea, which can be illustrated for a median: if the distribution of the variable of interest is such that the mean equals the median, the median could be estimated by using an estimator for the mean. We propose to transform the variable of interest in such a way that the median equals the mean for the transformed variable. If the transformation is monotone increasing, the median of the variable of interest can be estimated by inverting the estimate for the mean of the transformed variable. This method can also be extended to the estimation of any quantile. The proposed estimators are given by (18) and (20) in Subsection 3.3. In order to justify

this approach, it is necessary to transform the variable (Subsection 3.1) and to use a regression estimator (Subsection 3.2).

3.1. A Transformation of the Variables

We propose to transform the variable of interest such that the distribution of the transformed variable is approximately symmetric. Consider the midpoint distribution function $F^{\circ}(\cdot)$ (Nygård and Sandström 1985) defined by

$$F^{\circ}(y) = \frac{1}{2}[F(y^{-}) + F(y)]. \tag{10}$$

The quantity $F(y^{-})$ is the left-hand limit, that is, $F(y^{-}) = \lim_{t \to y^{-}} F(t)$. Alternatively, $F^{\circ}(y) = N^{-1} \sum_{i \in U} [\delta(y_i < y) + 0.5\delta(y_i = y)]$. Note that $0 < F^{\circ}(y_i) < 1$ for all $i \in U$. If the population quantile Y_α is the parameter of interest, we consider the following transformed values

$$y^{*}_{\alpha;i} = \Psi(y_i) + z_\kappa, \tag{11}$$

where $\Psi(y_i) = \phi^{-1}(F^{\circ}(y_i))$ and $\phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function $\phi(\cdot)$ of a normal $N(0, 1)$; that is,

$$\phi(y) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{y} \exp\left(\frac{-t^2}{2}\right) dt.$$

The quantity $z_\kappa = \phi^{-1}(\kappa)$ is the κ-th quantile of a normal $N(0, 1)$ distribution, with $\kappa = (\lceil \alpha N \rceil - 0.5)/N$. Note that κ can be approximated by α for large populations, as $\kappa \to \alpha$ when $N \to \infty$. The quantity α is the level of the quantile Y_α considered.

In the definition of $\Psi(y_i)$, we use (10) instead of $F(t)$ because the function $\phi^{-1}(\cdot)$ is not defined on 0 and 1. Note that the transformation $\Psi(y_i)$ does not depend on the choice of α. This function maps the quantiles of the distribution of y with the quantile of the standardised normal distribution $N(0, 1)$. Note that $\Psi(y_i)$ can be estimated with or without auxiliary variables.

The following Lemma gives the relationship between the population quantile Y_α and the following population mean of the transformed variable

$$\overline{Y}^{*}_{\alpha} = \frac{1}{N} \sum_{i \in U} y^{*}_{\alpha;i}.$$

Lemma 1 *We have that* $Y_\alpha = \Psi^{-1}(\overline{Y}^{*}_{\alpha})$, *where the function* $\Psi^{-1}(\cdot)$ *is the inverse of function* $\Psi(\cdot)$ *defined in (11)*

The proof is given in Appendix A.

The transformed values in (11) depend on population values, which would need to be estimated. We propose to estimate $y^{*}_{\alpha;i}$ by its substitution estimator given by

$$\widehat{y}^{*}_{\alpha;i} = \widehat{\Psi}(y_i) + z_\kappa,$$

where $\widehat{\Psi}(y_i) = \phi^{-1}(\widehat{F}^{\circ}(y_i))$. The function $\widehat{F}^{\circ}(\cdot)$ is the empirical midpoint estimator of the distribution function (10). This estimator is given by

$$\widehat{F}^{\circ}(y) = \frac{1}{2}[\widehat{F}(y^{-}) + \widehat{F}(y)], \tag{12}$$

where $\widehat{F}(\cdot)$ is a consistent estimator of $F(\cdot)$. In this article, we propose to use the Hájek-type estimator (3) in (12). However, we could use (6), (7) or (8) instead of (3). This may give a more efficient estimator.

The auxiliary variable may be transformed in the same way. When the values x_i are known for the entire population, we propose to use the following transformation.

$$x^{*}_{\alpha;i} = \Psi_x(x_i) + z_\kappa, \tag{13}$$

where $\Psi_x(x_i) = \phi^{-1}(F^{\circ}_x(x_i))$, $F^{\circ}_x(x) = [F_x(x^{-}) + F_x(x)]/2$ and $F_x(t) = N^{-1}\sum_{i\in U}\delta(x_i \le t)$. Note that the values of $x^{*}_{\alpha;i}$ cannot be calculated if we only know the sampled values of the auxiliary variable, as the function $F_x(\cdot)$ is unknown in this situation. If this is the case, we propose the transformation

$$\widehat{x}^{*}_{\alpha;i} = \widehat{\Psi}_x(x_i) + z_\kappa, \tag{14}$$

where $\widehat{\Psi}_x(x_i) = \phi^{-1}(\widehat{F}^{\circ}_x(x_i))$ and $\widehat{F}^{\circ}_x(x) = [\widehat{F}_x(x^{-}) + \widehat{F}_x(x)]/2$. The function $\widehat{F}_x(\cdot)$ may be any estimator of the distribution function $F_x(t)$. In this article, we propose to use the Hájek (1971) estimator of $F_x(\cdot)$ (see (3)).

3.2. *The Regression Estimator*

We propose to estimate $\overline{Y}^{*}_{\alpha}$ using a regression estimator (e.g. Cassel et al. 1976, 1977), which uses the auxiliary information. This estimator is defined by

$$\bar{y}^{*}_{reg;\alpha} = \bar{y}^{*}_{\alpha} + \widehat{\beta}_x\left(\overline{X}^{*}_{\alpha} - \bar{x}^{*}_{\alpha}\right), \tag{15}$$

where $\bar{y}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}\widehat{y}^{*}_{\alpha;i}$, $\overline{X}^{*}_{\alpha} = N^{-1}\sum_{i\in U}x^{*}_{\alpha;i}$, $\bar{x}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}x^{*}_{\alpha;i}$, with

$$\widehat{\beta}_x = \left[\sum_{i\in s}\frac{1}{\pi_i q_i^2}\left(x^{*}_{\alpha;i} - \bar{x}^{*}_{\alpha}\right)^2\right]^{-1}\sum_{i\in s}\frac{1}{\pi_i q_i^2}\left(x^{*}_{\alpha;i} - \bar{x}^{*}_{\alpha}\right)\left(\widehat{y}^{*}_{\alpha;i} - \bar{y}^{*}_{\alpha}\right). \tag{16}$$

Note that the regression estimator $\bar{y}^{*}_{reg;\alpha}$ assumes that the auxiliary variable is known for the entire population. When we only know the values of the auxiliary variable for the sampled units, we propose to use the following regression estimator instead of (15):

$$\bar{y}^{*}_{regS;\alpha} = \bar{y}^{*}_{\alpha} + \tilde{\beta}_x\left(\widehat{\overline{X}}^{*}_{\alpha} - \widehat{\bar{x}}^{*}_{\alpha}\right), \tag{17}$$

where $\widehat{\bar{x}}^{*}_{\alpha} = N^{-1}\sum_{i\in s}\pi_i^{-1}\widehat{x}^{*}_{\alpha;i}$ and $\widehat{\beta}_x$ is given by (16) after substituting $x^{*}_{\alpha;i}$ by $\widehat{x}^{*}_{\alpha;i}$. The control mean in (17) can be obtained as

$$\widehat{\overline{X}}^{*}_{\alpha} = \widehat{\Psi}_x(X_\alpha)\,.$$

This implicitly assumes that we know X_α. The Estimator (9) and the estimator proposed by Harms and Duchesne (2006) are also based on this assumption.

We can observe that the estimators $\bar{y}^*_{reg;\alpha}$ and $\bar{y}^*_{regS;\alpha}$ are based upon a single auxiliary variable. The proposed regression estimators can be easily extended to several auxiliary variables (e.g. Särndal et al. 1992, 225). For this purpose, the various auxiliary variables may be transformed by using the transformations (13) or (14) suggested for the variable x.

3.3. *The Proposed Estimators*

Based on Lemma 1, we propose to estimate the quantile Y_α by

$$\widehat{Y}_{reg;\alpha} = \widehat{\Psi}^{-1}\left(\bar{y}^*_{reg;\alpha}\right). \tag{18}$$

As $\widehat{\Psi}^{-1}(y) = \widehat{F}^{\circ -1}(\phi(y))$, an alternative expression for the proposed estimator is

$$\widehat{Y}_{reg;\alpha} = \widehat{F}^{\circ -1}(\widehat{\alpha}_{reg}), \tag{19}$$

where $\widehat{\alpha}_{reg} = \phi\left(\bar{y}^*_{reg;\alpha}\right)$. This estimator consists in inverting a midpoint distribution function $\widehat{F}^{\circ}(\cdot)$ at the value $\widehat{\alpha}_{reg}$, which is adjusted to take into account the auxiliary variable. Note that if we invert the midpoint distribution function (12) at the value α and if we use the estimator (3), we obtain an estimator which is approximately equal to the Hájek-type estimator (2) when $\widehat{F}^{\circ}(\cdot)$ is given by (3).

When we only know the values of the auxiliary variable for the sampled units and when the population quantile X_α is known, we propose to use a different estimator given by

$$\widehat{Y}_{regS;\alpha} = \widehat{\Psi}^{-1}\left(\bar{y}^*_{regS;\alpha}\right) = \widehat{F}^{\circ -1}(\widehat{\alpha}_{regS}), \tag{20}$$

where $\widehat{\alpha}_{regS} = \phi\left(\bar{y}^*_{regS;\alpha}\right)$ and $\bar{y}^*_{regS;\alpha}$ is defined by (17).

The proposed estimators are not affected by outliers, because $\widehat{y}^*_{\alpha;i}$ and $x^*_{\alpha;i}$ are implicitly based upon the ranks of y and x (see (11)). Note that $\widehat{Y}_{reg;\alpha} = X_\alpha$ when $y_i = x_i$. The efficiency of the proposed estimators depends on the correlation between y^*_i and x^*_i rather than the correlation between y_i and x_i.

It is worth investigating some properties of the Estimator (19) under equal probability sampling ($\pi_i = n/N$). In this case, it can be shown that

$$\bar{y}^*_{reg;\alpha} \doteq z_k - \widehat{\beta}_x \frac{1}{n}\sum_{i\in s}\Psi_x(x_i).$$

Thus, $\bar{y}^*_{reg;\alpha}$ increases monotonically when α increases, because z_κ is a monotone function of α, and $\widehat{\beta}_x$ and $\Psi_x(x_i)$ do not depend on α. Hence, $\widehat{Y}_{reg;\alpha_1} \leq \widehat{Y}_{reg;\alpha_2}$ when $\alpha_1 \leq \alpha_2$. This is a desirable property of an estimator of a quantile. Provided that $\widehat{\beta}_x > 0$, we have that $\widehat{\alpha}_{reg} > \alpha$ when $\sum_{i\in s}\Psi_x(x_i)$ is negative; that is, when the sample contains small x_i values. In this case, the estimate based on α (e.g. (2) with (3)) is likely to have a negative error. By using a level $\widehat{\alpha}_{reg}$ larger that α, we should reduce this error. Furthermore, as the adjustment, $\widehat{\beta}_x n^{-1}\sum_{i\in s}\Psi_x(x_i)$, does not depend on α, the proposed estimators are likely to be good for some α, but not for any α. The simulation study in Section 5 investigates this features.

The rescaled bootstrap variance estimator (Rao et al. 1992) can be used to estimate the variance of the proposed estimators. A confidence interval for the point estimator can be

also computed using the rescaled bootstrap confidence interval (the histogram approach). In Subsection 5.1, we evaluate the empirical performance of this variance estimator and this confidence interval.

4. Design Consistency

Consider the following regularity conditions:

$$|\widehat{Y}_\alpha - Y_\alpha| = O_p(n^{-1/2}), \tag{21}$$

$$\left|\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right| = O_p(n^{-1/2})\cdot \tag{22}$$

Conditions (21) and (22) mean that $\widehat{Y}_\alpha$ and $\bar{y}^*_{reg;\alpha}$ are $\sqrt{n}$-consistent. Isaki and Fuller (1982) and Robinson and Särndal (1983) gave conditions under which (22) holds. Francisco and Fuller (1991) established the consistency of $\widehat{Y}_\alpha$. Furthermore, the fact that the $y^*_{\alpha;i}$ can be considered as values generated from a normal distribution speaks in favour of (22).

As $\widehat{F}^{\circ -1}(\cdot)$ is a nondifferentiable function, we need to assume that this function converges to a differentiable function in order to prove the consistency. We assume that there exists a quantile function $Q(\cdot)$ which is twice differentiable, and such that

$$\sup_{|\epsilon|<o(n^{-1/2})} |\widehat{F}^{\circ -1}(\alpha+\epsilon) - \widehat{F}^{\circ -1}(\alpha) - Q(\alpha+\epsilon) + Q(\alpha)| = o_p(1). \tag{23}$$

This condition can be justified by Bahadur (1966) Lemma (see also Serfling 1980, Lemma E, p. 97).

Theorem 1 *Under assumptions (21), (22) and (23), the proposed estimator $\widehat{Y}_{reg;\alpha}$ is $\sqrt{n}$-consistent, as $|\widehat{Y}_{reg;\alpha} - Y_\alpha| = O_p(n^{-1/2})$.*

The proof of Theorem 1 is given in the Appendix B. In addition, $\widehat{Y}_{reg;\alpha}$ is asymptotically unbiased when $|\widehat{Y}_{reg;\alpha} - Y_\alpha|$ is uniformly bounded, as in this situation, the convergence in probability of $\widehat{Y}_{reg;\alpha}$ to Y_α implies that the expectation of $\widehat{Y}_{reg;\alpha}$ converges to Y_α (Lehmann 1999, 53).

It can be shown that the second estimator (20) is also consistent by assuming that (22) holds for $\bar{y}^*_{regS;\alpha}$.

5. Simulation

In this section, the proposed estimators $\widehat{Y}_{reg;\alpha}$ and $\widehat{Y}_{regS;\alpha}$ (see (18) and (20)) are compared numerically with alternative estimators described in Section 2. The alternative estimators considered are: $\widehat{Y}_{\pi;\alpha}$ (see (2)), $\widehat{Y}_{w;\alpha}$ (see (4)), $\widehat{Y}_{cd;\alpha}$ (Chambers and Dunstan 1986), $\widehat{Y}_{rkm;\alpha}$ (Rao et al. 1990), $\widehat{Y}_{ps;\alpha}$ (Silva and Skinner 1995), $\widehat{Y}_{r;\alpha}$ (see (9)) and $\widehat{Y}_{cal;\alpha}$ (Harms and Duchesne 2006).

The proposed Estimators (19) and (20) are based on the midpoint distribution function (12), which could be based on any estimator of $F(\cdot)$. For example, we can use the Estimators (3), (6), (7) or (8). The Estimators (6), (7) and (8) use auxiliary information and are therefore expected to be more accurate than (3). In our simulation study, we considered the worst-case scenario when the proposed estimators are based upon the Hájek-'type'

Table 1. Descriptive statistics of the variables of interest of the populations considered: ρ is the population correlation coefficient between y and x, ρ^* is the population correlation coefficient between y^* and x^*, and γ_y and γ_x are respectively the population skewness coefficients of y and x.

Pop.	$Y_{0.05}$	$Y_{0.25}$	$Y_{0.5}$	$Y_{0.75}$	$Y_{0.95}$	ρ	ρ^*	γ_y	γ_x
Sugar	34886	57585	80009	117159	204745	0.89	0.84	2.4	2.3
MUN-1	6	10	16	31	84	0.61	0.70	8.2	1.2
MUN-2	6	10	16	31	84	0.69	0.87	8.2	1.4
ES-SILC	13368	17970	22000	27700	42524	0.69	0.62	1.8	3.1
HMT	0.55	1.25	2.23	3.86	7.53	0.76	0.78	2.0	1.4

distribution function $\widehat{F}_\pi(t)$ defined by (3). In terms of simplicity, the proposed estimators should be obviously based upon (3).

The simulation study is based on several populations which are briefly described as follows. The sugar population consists of $N = 338$ sugar cane farms where y denotes the gross value of canes and x is the total cane harvested. The sugar population was used by Chambers and Dunstan (1986), Rao et al. (1990) and Silva and Skinner (1995). The population of municipalities (Särndal et al. 1992, 652) consists of $N = 284$ municipalities, where the variable of interest is the population size of the municipalities in 1985. We considered two auxiliary variables: (i) the number of conservative seats in municipal council (population MUN-1); and (ii) the total number of seats in municipal council (population MUN-2). We considered the Hansen et al. (1983) population (population HMT), which is $N = 14{,}000$ units generated from a bivariate gamma population (see also Rao et al. 1990). Finally, the last population is based on a random subset of $N = 2{,}000$ individuals from the 2012 Spanish Statistics on Income and Living Conditions (ES-SILC) Survey (Eurostat 2012). The ES-SILC provides information on income, poverty, social inclusion and living conditions for a sample of households and individuals. We considered the equivalised net income as the variable of interest and the tax on income contributions as the auxiliary variable. A brief descriptive analysis of the various populations is given in Table 1.

For each simulation, 1,000 samples were selected to compute the empirical relative bias $RB = (E[\widehat{Y}_\alpha] - Y_\alpha)/Y_\alpha$ and the empirical relative root mean square error $RRMSE = MSE[\widehat{Y}_\alpha]^{1/2}/Y_\alpha$ of an estimator $\widehat{Y}_\alpha$, where $E[\cdot]$ and $MSE[\cdot]$ denote respectively the empirical expectation and mean squared error. Simple random sampling and stratified random sampling were used to select the samples. The population quantiles $Y_{0.05}$, $Y_{0.25}$, $Y_{0.5}$, $Y_{0.75}$, and $Y_{0.95}$ are the parameters of interest.

Table 2 reports the empirical relative bias (RB) under simple random sampling. The RBs of the proposed estimators are of a reasonable range compared with the RBs of the alternative estimators, which can be larger than 10 percent in some cases. With the MUN-1 and MUN-2 populations, some estimators of $Y_{0.25}$ can have a large positive RB. Note that the proposed estimators tend to have large RB when the skewness of y is large and α is small or large. With $\alpha = 0.05$ or 0.95, the proposed estimators and the alternative estimators can have large positive RB, especially when $\alpha = 0.95$. For example, this is the case of the estimator $\widehat{Y}_{cal;\alpha}$ for the Sugar, MUN-1 and MUN-2 populations and when $\alpha = 0.95$. The simulation results indicate that the estimator $\widehat{Y}_{cal;\alpha}$ can be severely biased. The estimators $\widehat{Y}_{w;\alpha}$ and $\widehat{Y}_{\pi;\alpha}$ have similar RBs. Studies from the existing literature

Table 2. RB (%) of estimators of Y_α under simple random sampling.

Population	α	$\widehat{Y}_{\pi;\alpha}$	$\widehat{Y}_{w;\alpha}$	$\widehat{Y}_{cd;\alpha}$	$\widehat{Y}_{ps;\alpha}$	$\widehat{Y}_{rkm;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{r;\alpha}$	$\widehat{Y}_{cal;\alpha}$	$\widehat{Y}_{regS;\alpha}$
Sugar	0.05	0.1	0.9	−1.8	−1.3	−0.6	6.7	−5.1	2.8	4.4
($n = 30$)	0.25	0.1	−5.1	−0.5	0.3	−0.6	2.0	−1.2	1.7	1.8
	0.50	−2.1	−2.1	6.9	0.8	0.1	2.5	0.1	1.7	2.1
	0.75	−0.3	−6.0	10.7	0.3	0.1	2.5	−0.8	2.2	3.1
	0.95	3.2	1.5	3.0	4.6	−1.3	17.8	2.4	8.7	−2.7
Sugar	0.05	−5.3	−1.5	−8.1	−1.1	−1.6	2.5	−6.7	−0.2	0.6
($n = 60$)	0.25	−1.9	−2.0	−2.3	0.5	−0.4	0.9	−0.6	0.9	0.8
	0.50	−0.9	−1.0	4.3	1.0	0.1	1.3	0.8	1.1	1.4
	0.75	−1.8	−2.2	5.7	0.4	−0.4	0.8	1.3	0.8	1.0
	0.95	−1.9	1.7	10.8	2.6	2.6	6.8	2.7	9.0	7.1
MUN-1	0.05	−3.2	−3.2	−29.9	4.4	−3.7	6.6	0.3	3.4	8.8
($n = 50$)	0.25	5.1	−5.4	8.6	8.9	4.7	9.7	12.4	11.9	13.5
	0.50	−2.2	−6.5	31.3	2.5	−1.0	3.7	−4.6	1.5	−0.3
	0.75	0.3	−6.7	22.4	1.3	−0.4	3.4	−2.9	0.2	−0.3
	0.95	4.2	4.2	5.2	4.9	4.8	19.6	10.4	23.4	24.3
MUN-2	0.05	−2.9	−2.9	17.2	15.1	−3.3	7.3	−8.8	−2.8	−6.0
($n = 50$)	0.25	5.3	−5.4	21.9	17.3	4.8	9.8	5.4	23.5	17.1
	0.50	−1.9	−6.3	18.0	13.7	−0.4	4.1	−4.2	−0.2	−2.9
	0.75	0.1	−6.7	2.6	18.3	−1.2	1.9	0.5	23.1	10.5
	0.95	4.5	4.5	−6.7	28.9	4.5	19.8	5.5	20.5	21.4
HMT	0.05	−1.0	0.7	−53.9	0.4	−0.3	2.3	1.9	0.6	0.9
($n = 200$)	0.25	−0.3	0.3	−4.2	0.2	0.3	1.0	0.7	−0.3	0.9
	0.50	−0.1	0.4	10.4	0.4	0.4	0.9	0.4	−0.1	0.9
	0.75	−0.7	−0.1	11.2	0.0	0.1	0.4	0.0	−0.5	0.7
	0.95	−1.3	−1.2	10.0	0.4	0.6	2.1	0.0	−1.2	2.8
ES−SILC	0.05	−1.3	0.5	−8.9	0.3	0.4	2.2	−0.8	0.5	1.7
($n = 100$)	0.25	−0.3	−0.3	−3.5	0.1	0.1	0.4	−0.2	0.2	0.4
	0.50	−0.3	−0.3	0.8	0.1	0.0	0.4	−0.1	0.2	0.3
	0.75	−0.5	−0.5	5.3	−0.1	−0.1	0.4	0.4	0.3	0.4
	0.95	2.6	−0.3	11.1	−0.3	0.1	2.4	0.8	−3.0	2.3

(Dorfman 2009) indicate that the Chambers and Dunstan estimator, $\widehat{Y}_{cd;\alpha}$, can have a large bias. This estimator is based on a superpopulation model. Dorfman (2009) indicates that when the superpopulation model holds, this estimator tends to be very accurate. When the super population model does not hold, the estimator has an inevitable bias. This is the reason why we observe a large RBs for this estimator in Table 2. The large RB corresponds to situations when the superpopulation model does not hold.

The efficiency of the estimators is measured by the empirical relative root mean square errors (RRMSE) which are reported in Table 3. We observe that the proposed estimators perform well in all situations expect when $\alpha = 0.95$. However, we observe that the alternative estimators also have large RRMSE in this situation. Note that the proposed estimators are based upon the Hájek distribution function (3). We notice a clear

Table 3. RRMSE (%) of estimators of Y_α under simple random sampling.

Population	α	$\widehat{Y}_{\pi;\alpha}$	$\widehat{Y}_{w;\alpha}$	$\widehat{Y}_{cd;\alpha}$	$\widehat{Y}_{ps;\alpha}$	$\widehat{Y}_{rkm;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{r;\alpha}$	$\widehat{Y}_{cal;\alpha}$	$\widehat{Y}_{regS;\alpha}$
Sugar	0.05	17.7	17.7	15.0	18.2	16.4	18.1	16.6	17.8	18.6
($n = 30$)	0.25	11.6	12.5	6.6	9.7	9.2	9.3	10.7	9.4	9.4
	0.50	12.0	11.2	9.8	9.6	9.3	9.4	10.6	10.4	10.5
	0.75	14.3	13.1	15.3	10.9	10.3	11.4	11.2	12.6	12.5
	0.95	26.0	22.1	54.9	27.8	31.7	42.6	17.8	35.3	51.3
Sugar	0.05	13.8	13.1	12.9	12.5	12.2	12.1	14.0	12.6	12.9
($n = 60$)	0.25	8.2	8.0	4.6	6.2	6.2	6.3	7.7	6.2	6.4
	0.50	8.3	7.6	6.0	6.5	6.1	6.2	7.3	7.0	7.1
	0.75	8.9	7.7	7.3	6.4	5.9	6.6	7.0	6.9	6.8
	0.95	12.4	12.0	29.0	14.3	13.7	18.2	12.9	27.1	28.1
MUN-1	0.05	19.5	19.5	33.3	17.8	19.1	18.3	25.9	18.2	18.7
($n = 50$)	0.25	12.2	12.1	13.3	14.0	12.0	14.0	18.7	15.8	18.4
	0.50	14.8	15.5	34.7	15.2	13.3	13.3	14.1	14.4	13.0
	0.75	17.1	15.3	29.5	14.9	12.4	17.8	14.0	14.4	13.5
	0.95	29.6	29.4	55.7	33.8	38.7	52.7	29.2	92.4	92.2
MUN-2	0.05	18.6	18.6	21.4	22.5	18.2	17.3	18.2	19.4	16.8
($n = 50$)	0.25	12.7	12.7	23.8	23.0	11.1	13.8	12.9	25.7	19.1
	0.50	14.4	14.9	22.5	26.1	12.3	11.9	12.4	12.6	11.0
	0.75	16.7	16.3	15.4	26.0	13.2	12.1	13.1	32.6	15.3
	0.95	28.0	28.0	26.7	77.9	28.0	58.4	23.7	76.9	83.7
HMT	0.05	11.7	11.5	55.1	11.3	12.1	11.4	19.6	11.8	12.7
($n = 200$)	0.25	8.0	7.6	6.0	6.5	6.2	6.5	7.7	8.0	6.9
	0.50	7.5	6.8	11.2	5.9	5.8	5.9	6.4	7.5	6.3
	0.75	7.2	6.3	11.9	5.7	5.5	5.7	6.4	7.2	6.1
	0.95	9.9	9.1	11.9	9.6	8.9	9.9	9.3	9.9	11.0
ES-SILC	0.05	8.3	8.1	11.4	7.8	7.8	7.9	8.4	8.1	9.1
($n = 100$)	0.25	3.7	3.6	4.4	3.4	3.3	3.4	3.9	3.6	3.6
	0.50	4.0	3.8	2.7	3.3	3.3	3.3	3.8	3.6	3.6
	0.75	4.7	4.2	6.1	4.0	3.6	3.8	4.7	4.1	4.1
	0.95	10.6	10.1	18.7	10.4	10.2	11.3	11.0	10.5	12.8

improvement between the proposed estimators and the Hájek estimator (2), because the RRMSEs of the proposed estimators are usually smaller than the RRMSEs of the Hájek estimator $\widehat{Y}_{\pi;\alpha}$. In other words, there is a clear improvement when using $\widehat{\alpha}_{reg}$ instead of α, except when $\alpha = 0.95$ and 0.25 with the MUN-1 and MUN-2 populations. The proposed estimators can be more efficient than the alternative estimators, especially when $\alpha = 0.50$ and 0.75. We also observe that $\widehat{Y}_{reg;\alpha}$ is generally more efficient than $\widehat{Y}_{regS;\alpha}$.

We also conducted another series of simulations using stratified simple random sampling. The conclusions derived from this simulation study are similar. The results of this simulation study are not presented in this article.

We now investigate the conditional relative biases of the proposed estimator $\widehat{Y}_{reg;\alpha}$ given the sample means of the auxiliary variable. For this purpose, the 1,000 selected samples were ordered according to the mean of the auxiliary variable. Then this ranking was used to create 20 groups of 50 observations each. Conditional relative biases were then obtained by calculating the *RB* for each of the 20 groups.

Figure 1 displays the conditional relative biases of the estimators of the first quartile under simple random sampling from the Sugar population. We observe that the Hájek-type estimator clearly exhibits the worst conditional performance with a linear trend as the group mean of x increases. The conditional RB of the proposed estimator and the Rao et al. (1990) estimator does not seem to be correlated with the group mean of x. The Rao et al. (1990) estimator has a bias which is slightly smaller than the bias of the proposed estimator. Figure 2 displays the conditional relative biases of the estimators of the median under simple random sampling from the MUN-1 population. The conditional relative bias of the proposed estimator and the Rao et al. (1990) estimator does not seem to be correlated with the group mean of x.

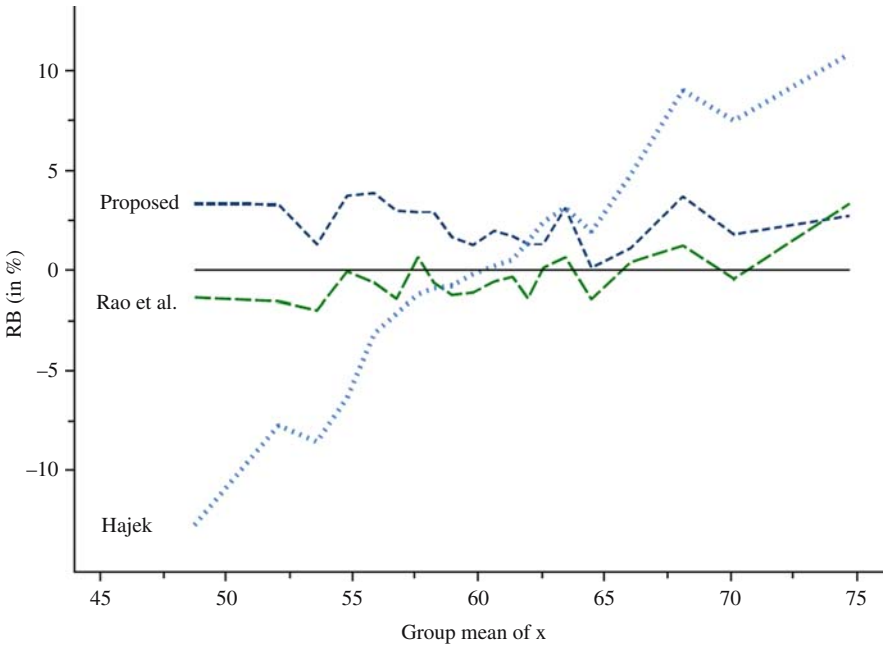

Fig. 1. Conditional relative biases (%) of estimates of $Y_{0.25}$ under simple random sampling from the sugar population when $n = 30$.

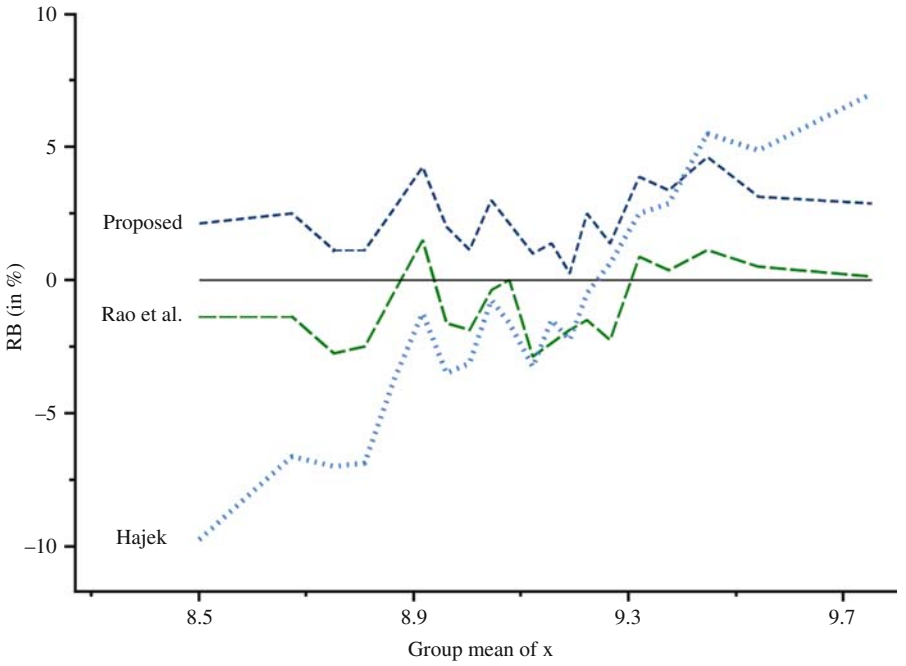

Fig. 2. Conditional relative biases (%) of estimates of $Y_{0.5}$ under simple random sampling from the MUN-1 population when $n = 200$.

The proposed estimator is biased and $\widehat{Y}_{rkm;\alpha}$ is approximately unbiased. This explains why $\widehat{Y}_{rkm;\alpha}$ shows under- and overestimation in Figures 1 and 2, otherwise $\widehat{Y}_{rkm;\alpha}$ would not be approximately unbiased. We observe an overestimation for all groups of mean for the proposed estimator, because this estimator has a small non-negligible bias.

The proposed transformation-based approach seems to perform well for estimating the central quantiles. In particular, results derived from simulation studies indicate that the proposed estimators have a good performance for the median. In this situation, the proposed estimators clearly outperform the Hájek estimator, especially when the conditional bias is taken into consideration. In addition, the proposed estimators perform well if they are compared to the various existing methods. For instance, although the proposed estimators can be slightly biased, they seem more efficient than the simpler alternatives $\widehat{Y}_{r;\alpha}$ (the ratio estimator) and $\widehat{Y}_{cal;\alpha}$ (Harms and Duchesne 2006). The values of RRMSE of the proposed estimators are comparable to the values of RRMSE of the more sophisticated estimator $\widehat{Y}_{rkm;\alpha}$ (Rao et al. 1990). These conclusions hold also in the situation where only population quantiles of the auxiliary variable are known. However, the proposed estimators can have large biases for the tail quantiles, specially when $\alpha = 0.95$. In this situation, the Hájek estimator appears more robust compared to all the more complex approaches.

5.1. Variance Estimation and Confidence Intervals

We propose to estimate the variance of the proposed point estimators using the rescaled bootstrap variance estimator (Rao et al. 1992). Rao and Wu (1988) showed that the rescaled bootstrap variance estimator is a consistent estimator for the variance when the

Table 4. Empirical relative bias (%) of the rescaled bootstrap variance estimators under simple random sampling when $n = 200$. The column ρ gives the correlation between the auxiliary variable and the variable of interest.

Population	ρ	$\frac{n}{N}$	$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$	
			$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$
ES-SILC	0.69	0.01	8.0	7.3	12.6	9.5	18.0	18.9
		0.05	13.3	11.7	24.4	23.8	14.8	11.9
Log-Normal	0.50	0.01	13.1	11.0	5.6	5.4	6.1	0.6
		0.05	23.0	17.6	18.4	16.8	12.9	10.5
	0.70	0.01	2.1	6.7	14.4	12.0	8.4	6.5
		0.05	17.5	10.5	15.1	13.3	14.2	18.2
	0.90	0.01	4.4	10.7	3.4	8.1	17.8	12.6
		0.05	22.4	20.6	17.4	19.8	28.9	24.5
HMT	0.76	0.014	8.0	16.9	7.4	4.1	7.5	9.5

sampling fraction is small. A confidence interval can be computed using the rescaled bootstrap confidence interval (the histogram approach). In this section, we evaluate the empirical performance of this variance estimator and this confidence interval. A set of 10,000 independent simple random samples were selected.

We used the ES-SILC and HMT populations defined in Section 5. In addition, we used artificial populations with variables of interest generated from log-normal distributions. Auxiliary variables correlated with the variable of interest are randomly generated. We consider the following correlation coefficients: 0.5, 0.7 and 0.9. The sample size considered is $n = 200$. The sampling fractions considered are $n/N = 0.01, 0.014$ and 0.05.

In Table 4, we have the empirical relative biases of the rescaled bootstrap variance estimator. We observe larger relative biases when the sampling fraction is 0.05. The bias does not seem to be affected by the correlation or the level α. In Table 5, we have the

Table 5. Coverage rates (%) of the 95 percent rescaled bootstrap confidence interval (the histogram approach) under simple random sampling when $n = 200$.The column ρ gives the correlation between the auxiliary variable and the variable of interest.

Population	ρ	$\frac{n}{N}$	$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$	
			$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$	$\widehat{Y}_{reg;\alpha}$	$\widehat{Y}_{regS;\alpha}$
ES-SILC	0.69	0.01	94.6	94.7	93.8	93.7	94.5	93.9
		0.05	94.6	94.9	95.8	95.9	95.1	95.2
Log-Normal	0.50	0.01	96.0	95.7	94.7	94.4	93.7	93.4
		0.05	96.2	96.4	95.7	95.6	96.3	96.4
	0.70	0.01	95.7	96.4	96.9	96.1	95.3	94.7
		0.05	97.2	97.1	96.2	94.9	95.2	95.5
	0.90	0.01	95.4	96.0	94.2	94.9	96.4	95.7
		0.05	95.7	95.5	96.0	96.4	95.3	95.8
HMT	0.76	0.014	93.3	94.8	94.3	93.5	94.6	94.3

observed coverage rates of the 95 percent rescaled bootstrap confidence interval. All the coverages observed are close to the nominal level of 95 percent. Based on this limited simulation study, it seems preferable to consider bootstrap confidence intervals rather than bootstrap variance, when measuring the accuracy of the proposed estimators.

6. Discussion

The proposed estimators are based on a regression estimator of the population mean, which is a technique widely used with survey data. The proposed approach can be applied to many standard surveys. It can be implemented with multistage sampling designs, as the proposed estimators are based upon first-order inclusion probabilities and a regression estimator. Alternative estimators proposed by Chambers and Dunstan (1986) and Rao et al. (1990) can be slightly more accurate than the proposed estimators. However, in order to compute these alternative estimators, it is necessary to know the auxiliary variable for the entire population. The Rao et al. (1990) estimator also requires the joint inclusion probabilities, which can be unknown. The proposed estimators are computationally simpler because they are free of joint inclusion probabilities, they are based on a regression estimator and they can be computed when the auxiliary variable is unknown for the nonsampled units. When the joint inclusion probabilities are known, the accuracy of the proposed estimators can also be improved by inverting the Rao et al. (1990) estimator of the distribution function (or any other estimators) rather than the Hájek-type estimator of the distribution function.

We have considered a regression estimator to take the auxiliary information into account. Other type of estimators based upon auxiliary information (Huang and Fuller 1978:, Deville and Särndal 1992) can also be used instead of a regression estimator. The proposed estimators can also be generalised to several auxiliary variables, since a regression estimator can be easily extended to accommodate this situation. In this article, the auxiliary variables are used to calibrate toward a population mean. This approach can be extended to calibration towards more complex population quantities such as means, quantiles, or variances (e.g. Owen 1991, Chaudhuri et al. 2008, Lesage 2011).

Chen and Wu (2002) proposed a pseudoempirical likelihood approach for estimating quantiles with auxiliary variables. Berger and De la Riva Torres (2015) proposed an empirical-likelihood approach for estimating quantiles with auxiliary variables. Empirical (and pseudoempirical) likelihood approaches are well suited for the estimation of quantiles with auxiliary variables, especially for the calculation of confidence intervals. It would be interesting to investigate how an empirical-likelihood approach could be used to derived confidence intervals for the proposed approach.

Appendix A: Proof of Lemma 1

We have that

$$\overline{Y}^*_\alpha = \frac{1}{N}\sum_{i\in U} y^*_{\alpha;i} = \frac{1}{N}\sum_{i\in U} \phi^{-1}(F^\circ(y_i)) + z_\kappa, \tag{24}$$

$$F^\circ(y_i) = R_i, \tag{25}$$

where $R_i = N^{-1}(rank(y_i) - 0.5)$ and $rank(y_i)$ is the rank of observation y_i in the population and $\phi^{-1}(\cdot)$ is the quantile function of a $N(0, 1)$ distribution. By substituting (25) into (24), we have that

$$\overline{Y}_{\alpha}^{*} = \frac{1}{N}\sum_{i\in U}\phi^{-1}(R_i) + z_{\kappa} = \frac{1}{N}(S_{<0.5} + S_{>0.5} + S_{0.5}) + z_{\kappa} \tag{26}$$

with

$$S_{<0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i < 0.5),$$

$$S_{>0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i > 0.5),$$

$$S_{0.5} = \sum_{i\in U}\phi^{-1}(R_i)\delta(R_i = 0.5).$$

It is clear that $S_{0.5} = 0$. Consider a unit i such that $rank(y_i) < (N + 1)/2$. This implies that $R_i < 0.5$. Thus

$$S_{<0.5} = \sum_{r<(N+1)/2}\phi^{-1}((r - 0.5)/N), \tag{27}$$

$$\begin{aligned} S_{>0.5} &= \sum_{r<(N+1)/2}\phi^{-1}((N - r + 1 - 0.5)/N) \\ &= \sum_{r<(N+1)/2}\phi^{-1}(1 - (r - 0.5)/N). \end{aligned} \tag{28}$$

Substituting (27) and (28) into (26), we obtain

$$\overline{Y}_{\alpha}^{*} = \frac{1}{N}\sum_{r<(N+1)/2}\left\{\phi^{-1}((r - 0.5)/N) + \phi^{-1}(1 - (r - 0.5)/N)\right\} + z_{\kappa}. \tag{29}$$

As the normal distribution is symmetric, we have that $\phi^{-1}(p) = -\phi^{-1}(1 - p)$. Hence the sum in (29) equal zero. This implies that

$$\overline{Y}_{\alpha}^{*} = z_{\kappa}. \tag{30}$$

As $F^{\circ}(Y_{\alpha}) = N^{-1}(rank(Y_{\alpha}) - 0.5)$, $rank(Y_{\alpha}) = \lceil \alpha N \rceil$, and $\kappa = N^{-1}(\lceil \alpha N \rceil - 0.5)$, we have that

$$F^{\circ}(Y_{\alpha}) = \kappa. \tag{31}$$

We also have that

$$F^{\circ}(Y_{\alpha}) = \phi(\phi^{-1}(F^{\circ}(Y_{\alpha}))) = \phi(\Psi(Y_{\alpha})). \tag{32}$$

Equations (31) and (32) imply that

$$\phi(\Psi(Y_{\alpha})) = \kappa. \tag{33}$$

As z_κ is the κth quantile of a normal $N(0, 1)$ distribution, we have that $\phi(z_\kappa) = \kappa$, which combined with (33) gives

$$\phi(z_\kappa) = \phi(\Psi(Y_\alpha)) \cdot$$

The last expression implies

$$z_\kappa = \Psi(Y_\alpha), \tag{34}$$

as $\phi(\cdot)$ is a bijective function. Combining (30) with (34), we have that $\Psi(Y_\alpha) = \overline{Y}^*_\alpha$. The Lemma follows.

Appendix B: Proof of Theorem 1

As $\phi(\cdot)$ is twice differentiable, a first-order Taylor expansion implies that

$$\phi\left(\bar{y}^*_{reg;\alpha}\right) - \phi\left(\overline{Y}^*_\alpha\right) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) + O_p\left(\left|\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right|^2\right), \tag{35}$$

where $f(y)$ is the density of a $N(0, 1)$ distribution. Equation (30) implies that $\phi\left(\overline{Y}^*_\alpha\right) = \phi(z_\kappa) = \kappa$. Thus, as $\kappa \to \alpha$ as $N \to \infty$, $\lim_{N\to\infty} \phi\left(\overline{Y}^*_\alpha\right) = \alpha$ and we have that

$$\phi(\bar{y}^*_{reg;\alpha}) - \alpha = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) + O_p(n^{-1}), \tag{36}$$

because $\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha = O_p(n^{-1/2})$.

As $Q(\alpha)$ is twice differentiable, a first-order Taylor expansion implies that

$$Q\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - Q(\alpha) = \left(\phi\left(\bar{y}^*_{reg;\alpha}\right) - \alpha\right) Q'(\alpha) + O_p\left(\left|\phi\left(\bar{y}^*_{reg;\alpha}\right) - \alpha\right|^2\right),$$

where $Q'(\alpha) = \partial Q(\alpha)/\partial\alpha$. Assumption (22) and (36) imply that

$$Q\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - Q(\alpha) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}), \tag{37}$$

as $f\left(\overline{Y}^*_\alpha\right)$ is bounded. Using assumption (23), Equation (37) implies that

$$\widehat{F}^{\circ -1}\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) - \widehat{F}^{\circ -1}(\alpha) = \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}). \tag{38}$$

As $\widehat{F}^{\circ -1}\left(\phi\left(\bar{y}^*_{reg;\alpha}\right)\right) = \widehat{Y}_{reg;\alpha}$ and $\widehat{F}^{\circ -1}(\alpha) = \widehat{Y}_\alpha$, equation (38) becomes

$$\widehat{Y}_{reg;\alpha} = \widehat{Y}_\alpha + \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1})$$

which implies

$$\widehat{Y}_{reg;\alpha} - Y_\alpha = \widehat{Y}_\alpha - Y_\alpha + \left(\bar{y}^*_{reg;\alpha} - \overline{Y}^*_\alpha\right) f\left(\overline{Y}^*_\alpha\right) Q'(\alpha) + O_p(n^{-1}).$$

Thus, the last expression combined with the conditions (21) and (22) implies that $|\widehat{Y}_{reg;\alpha} - Y_\alpha| = O_p(n^{-1/2})$.

7. References

Bahadur, R.R. 1966. "A Note on Quantiles in Large Samples." *The Annals of Mathematical Statistics* 37: 577–580.

Berger, Y.G. and C.J. Skinner. 2003. "Variance Estimation of a Low-Income Proportion." *Journal of the Royal Statistical Society Series C* 52: 457–468. DOI: http://dx.doi.org/10.1111/1467-9876.00417.

Berger, Y.G. and O. De la Riva Torres. 2015. "An Empirical Likelihood Approach for Inference Under Complex Sampling Design. To Appear in Journal of Royal Statistical Society, Senes B, 22p.".

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1976. "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations." *Biometrika* 63: 615–620. DOI: http://dx.doi.org/10.1093/biomet/63.3.615.

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1977. *Foundation of Inference in Survey Sampling*. New York: Wiley.

Chambers, R.L. and R. Dunstan. 1986. "Estimating Distribution Functions From Survey Data." *Biometrika* 73: 597–604. DOI: http://dx.doi.org/10.1093/biomet/73.3.597.

Chaudhuri, S., M.S. Handcock, and M.S. Rendall. 2008. "Generalized Linear Models Incorporating Population Level Information: An Empirical-Likelihood-Based Approach." *Journal of the Royal Statistical Society – Series B (Statistical Methodology)* 70: 311–328. DOI: http://dx.doi.org/10.1111/j.1467-9868.2007.00637.x.

Chen, J. and C. Wu. 2002. "Estimation of Distribution Function and Quantiles Using Model-Calibrated Pseudo Empirical Likelihood Method." *Statistica Sinica* 12: 1223–1239.

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: http://dx.doi.org/10.1080/01621459.1992.10475217.

Dorfman, A.H. 2009. "Inference on Distribution Functions and Quantiles." In *Handbook of Statistics 29B Sample Surveys: Inference and Analysis*, edited by D. Pfeffermann and C.R. Rao, pp. 371–395. Amsterdam, North-Holland: Elsevier.

Eurostat. 2003. *"Laeken" Indicators-Detailed Calculation Methodology, Directorate E: Social Statistics, Unit E-2: Living Conditions, DOC.E2/IPSE/2003*. Available at: http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken%20Indicators%20-%20calculation%20algorithm.pdf.

Eurostat. 2012. *European Union Statistics on Income and Living Conditions (EU-SILC)*. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc.

Francisco, C.A. and W.A. Fuller. 1991. "Quantile Estimation With a Complex Survey Design." *Annals of Statistics* 19: 454–469.

Hájek, J. 1971. Comment on a paper by D. Basu. In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston.

Hansen, M.H., W.G. Madow, and B.J. Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78: 776–793. DOI: http://dx.doi.org/10.1080/01621459.1983.10477018.

Harms, T. and P. Duchesne. 2006. "On Calibration Estimation for Quantiles." *Survey Methodology* 32: 37–52.

Huang, E.T. and W.A. Fuller. 1978. "Nonnegative Regression Estimation for Survey Data." In Procceding of the Social Statistics Section of the American Statistical Association, Washington DC, 300–303.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. DOI: http://dx.doi.org/10.1080/01621459.1982.10477770.

Lehmann, E.L. 1999. *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Lesage, E. 2011. "The Use of Estimating Equations to Perform a Calibration on Complex Parameters." *Survey Methodology* 37: 103–108.

Nygård, F. and A. Sandström. 1985. "The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations." *Journal of Official Statistics* 4: 399–412.

Osier, G. 2009. "Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques." *Journal of the European Survey Research Association* 3: 167–195.

Owen, A.B. 1991. "Empirical Likelihood for Linear Models." *The Annals of Statistics* 19: 1725–1747.

Rao, J.N.K., J.G. Kovar, and H.J. Mantel. 1990. "On Estimating Distribution Functions and Quantiles From Survey Data Using Auxiliary Information." *Biometrika* 77: 365–375. DOI: http://dx.doi.org/10.1093/biomet/77.2.365.

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83: 231–241. DOI: http://dx.doi.org/10.1080/01621459.1988.10478591.

Rao, J.N.K., C.F.J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18: 209–217.

Robinson, P.M. and C.-E. Särndal. 1983. "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling." *Sankhya B* 45: 240–248.

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

Serfling, N. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Silva, P.L.D., Nascimento and C.J. Skinner. 1995. "Estimating Distribution Functions With Auxiliary Information Using Poststratification." *Journal of Official Statistics* 11: 277–294.

Verma, V. and G. Betti. 2011. "Taylor Linearization Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics." *Journal of Applied Statistics* 38: 1549–1576. DOI: http://dx.doi.org/10.1080/02664763.2010.515674.

Received October 2013
Revised October 2014
Accepted November 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 121–138, http://dx.doi.org/10.1515/JOS-2015-0006

Statistical Disclosure Limitation in the Presence of Edit Rules

Hang J. Kim[1], *Alan F. Karr*[2], *and Jerome P. Reiter*[3]

We compare two general strategies for performing statistical disclosure limitation (SDL) for continuous microdata subject to edit rules. In the first, existing SDL methods are applied, and any constraint-violating values they produce are replaced using a constraint-preserving imputation procedure. In the second, the SDL methods are modified to prevent them from generating violations. We present a simulation study, based on data from the Colombian Annual Manufacturing Survey, that evaluates the performance of the two strategies as applied to several SDL methods. The results suggest that differences in risk-utility profiles across SDL methods dwarf differences between the two general strategies. Among the SDL strategies, variants of microaggregation and partially synthetic data offer the most attractive risk-utility profiles.

Key words: Confidentiality; imputation; survey; synthetic data.

1. Introduction

Public-use microdata offer many benefits, for example, enabling researchers and policy makers to perform in-depth statistical analyses, students to learn skills in data analysis, and citizens to understand their society. However, public-use microdata also carry disclosure risks: intruders who intend to misuse the information may be able to identify respondents or learn values of sensitive attributes from the public data. Statistical agencies recognize this risk and typically alter the microdata prior to release using one or more statistical disclosure limitation (SDL) techniques. Ideally, the SDL reduces disclosure risk to an acceptable level with low impact on data utility (Willenborg and De Waal 2001; Hundepool et al. 2012).

As collected, microdata often include implausible or impossible values, for example arising from multiple forms of survey error (Groves 1989) such as reporting and measurement error. Agencies prefer not to release such faulty values and so undertake a process usually referred to as "edit and imputation" (De Waal et al. 2011). Agencies identify faulty values via prespecified constraints, called *edit rules* or simply *edits*.

[1] Duke University and National Institute of Statistical Sciences, P.O. Box 90251, Durham, NC 27708, U.S.A. Email: hangkim0@gmail.com
[2] RTI International, 3040 East Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709, U.S.A. Email: karr@rti.org
[3] Department of Statistical Science, Duke University, P.O. Box 90251, Durham, NC 27708, U.S.A. Email: jerry@stat.duke.edu
Acknowledgments: This research was supported by the National Science Foundation (SES-11-31897). The authors thank the Editor, the Associate Editor, and the three anonymous referees for their insightful and constructive comments.

Examples of edit rules for continuous microdata, such as data from economic censuses or surveys, include *range restrictions* ($V_1 \leq a$), *ratio constraints* ($V_1 \leq bV_2$), and *balance constraints* ($V_1 + V_2 = V_3$). When a record fails a set of edits, agencies typically select some fields to replace with imputed values so that all constraints are satisfied (Fellegi and Holt 1976).

To date, assessment of disclosure risks and subsequent SDL have been largely disconnected from edit and imputation in practice. Typically editing is performed by one organizational unit, which then transfers the data to another unit that performs SDL. Interaction between the editing and SDL processes is minimal, and sometimes is entirely absent. Indeed, those performing the SDL may not even be aware of constraints that the edited data must respect.

The extant literature offers two general strategies for integrating SDL and editing. The first approach is to apply existing SDL methods and then remove any resulting edit violations; this is illustrated in Shlomo and De Waal (2005; 2008). Essentially, edit violations engendered by SDL are treated in the same way as those resulting from measurement error. The second approach is to use an SDL method that does not produce edit violations; this is illustrated in Torra (2008). Many SDL methods as typically applied do not guarantee edit preservation; however, as we illustrate, some SDL methods can be modified to do so. To our knowledge, these two general strategies have not been compared in terms of impacts on data quality and disclosure risk.

In this article, we make such comparisons by implementing the strategies for several SDL procedures for continuous microdata. We apply the procedures to continuous microdata from the 1991 Colombian Annual Manufacturing Survey. The results of the simulation suggest that, when both strategies are feasible, there is little difference in the risk-utility profiles of edit-after-SDL (first approach) and edit-preserving SDL (second approach) procedures. Indeed, the differences in the profiles across approaches are swamped by differences among SDL methods. We also discuss the relative merits of the SDL techniques, although we view the evidence from the simulations as more suggestive than complete.

The remainder of the article is organized as follows. In Section 2, we describe several SDL methods and corresponding approaches to generate masked values satisfying edits. In Section 3, we present results of the simulation study and compare the suggested methods under a risk-utility framework. In Section 4, we conclude with a discussion of future research questions.

2. SDL Methods in the Presence of Edit Rules

As in Reiter (2005), let y_{il} be the collected value of variable l for unit i, for $l = 0, \ldots, p$ and $i \in D$, where D denotes the collected data for the n sampled units. Let y_{i0} be the unique unit identifier, which, if it is informative, must be excluded from the final released data. Suppose that $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{ip}\}$ satisfies all constraints or has been corrected to do so prior to SDL. For each $i \in D$, let $\boldsymbol{y}_i$ be partitioned as $(\boldsymbol{y}_i^A, \boldsymbol{y}_i^U)$, where $\boldsymbol{y}_i^A$ is a vector of variables available to intruders in external data files, and $\boldsymbol{y}_i^U$ is a vector of variables unavailable to intruders except in the released data file, D^{rel}. To prevent disclosure, the agency uses SDL to alter the values of $\boldsymbol{y}_i^A$ before releasing D^{rel}. Let $\tilde{\boldsymbol{y}}_i^A$ denote the masked

values of $\boldsymbol{y}_i^A$, so that D^{rel} after SDL comprises $\tilde{\boldsymbol{y}}_i = (\tilde{\boldsymbol{y}}_i^A, \boldsymbol{y}_i^U)$ for all n records on the file. For simplicity, we assume that the intruder knows $\boldsymbol{y}_i^A$ without any measurement error. In general, it is challenging for agencies to determine which variables comprise $\boldsymbol{y}_i^A$ and which comprise $\boldsymbol{y}_i^U$. When this distinction is unclear, arguably the agency should treat all variables as needing disclosure treatment.

2.1. *Summary of Selected SDL Methods*

In this section, we review the set of SDL methods for continuous microdata that we employ in our simulation, which includes rank swapping, adding noise, variants of microaggregation, and partially synthetic data. We describe each method briefly and refer readers to Hundepool et al. (2012) for further details. Of course, there are more variations on these methods, as well as additional SDL methods. We do not claim that these are a subset of best or most appropriate methods for the data at hand; however, they do serve to help us evaluate the two general strategies for SDL with editing.

Rank swapping (Moore 1996) is a special form of data swapping under which some attribute values are switched between pairs of similar records. Rank swapping is implemented as follows. For each variable l in $\boldsymbol{y}_i^A$, we sort $\{y_{1l}, \ldots, y_{nl}\}$ by its magnitude; let $\{y_{(1)l}, \ldots, y_{(n)l}\}$ denote the ordered values. Let $0 < \tau_{\text{swap}} < 100$ be a prespecified parameter. Two cases $y_{(i)l}$ and $y_{(j)l}$ are randomly selected, and then swapped only if $|i - j| < n\tau_{\text{swap}}/100$. As τ_{swap} increases, the intensity of data protection increases but, in general, the data utility decreases.

Adding noise (Kim 1986; Sullivan and Fuller 1990; Tendick 1991) introduces random errors to selected values deemed at high risk of disclosure; for example, set $\tilde{\boldsymbol{y}}_i^A = \boldsymbol{y}_i^A + \boldsymbol{\varepsilon}_i$. A straightforward implementation is to draw random noise from a normal distribution, $\boldsymbol{\varepsilon}_i \sim N(0, \tau_{\text{noise}}\Sigma^A)$, where Σ^A is the sample covariance of $\{\boldsymbol{y}_1^A, \ldots, \boldsymbol{y}_n^A\}$. The agency sets the parameter τ_{noise} to control the intensity of perturbation. To increase data utility, Shlomo and de Waal (2008) suggest perturbing data within control strata, in which the agency (i) defines Q subgroups of records $\{D_q : q = 1, \ldots, Q\}$, for example, by grouping records into quintiles of some variable, (ii) generates random noise $\boldsymbol{\varepsilon}_i \sim N\left(\boldsymbol{\mu}_q\left(1 - \sqrt{1 - \tau_{\text{cnoise}}^2}\right)/\tau_{\text{cnoise}}, \Sigma_q\right)$ where $\boldsymbol{\mu}_q$ and Σ_q are the sample mean and the sample covariance of records $\{\boldsymbol{y}_j^A : j \in D_q\}$ and $0 < \tau_{\text{cnoise}} \leq 1$ is the parameter to control the amount of random noise, and (iii) replaces $\boldsymbol{y}_i^A$ with $\tilde{\boldsymbol{y}}_i^A = \sqrt{1 - \tau_{\text{cnoise}}^2}\boldsymbol{y}_i^A + \tau_{\text{cnoise}}\boldsymbol{\varepsilon}_i$. We refer to this variation as *controlled adding noise*.

Microaggregation (Defays and Nanopoulos 1993; Domingo-Ferrer and Mateo-Sanz 2002) replaces original values with group averages. Using a clustering algorithm, the original records $\boldsymbol{y}_i$ are partitioned into clusters $\mathcal{G}_g$, each with a fixed size. For each $i \in \mathcal{G}_g$, we replace $\boldsymbol{y}_i^A$ with the group mean $\tilde{\boldsymbol{y}}_{\text{mic},i}^A = \sum_{k \in \mathcal{G}_g} \boldsymbol{y}_k^A / \tau_{\text{mic}}$, where $\tau_{\text{mic}} = |\mathcal{G}_g|$, the cardinality of $\mathcal{G}_g$. Larger cluster sizes result in greater data perturbation. To construct clusters, one can project data onto a single dimension, for example, using the first principal component or the sum of z-scores (Fayyoumi and Oommen 2010). Alternatively, one can find the clusters using a heuristic based on Euclidean distances between records. For example, in *multivariate fixed-size microaggregation* (Domingo-Ferrer and

Mateo-Sanz 2002), the algorithm starts with finding the two records $\boldsymbol{y}_r$ and $\boldsymbol{y}_s$ farthest apart. The first cluster contains $\boldsymbol{y}_r$ and the $\tau_{\text{mmic}} - 1$ records closest to $\boldsymbol{y}_r$, and the second cluster contains $\boldsymbol{y}_s$ and the $\tau_{\text{mmic}} - 1$ records closest to $\boldsymbol{y}_s$. The third and fourth clusters are formed in a similar fashion starting from the two farthest-apart records among the remaining $n - 2\tau_{\text{mmic}}$ records. This repeats until fewer than $2\tau_{\text{mmic}}$ records do not belong to the clusters. These remaining records form a new cluster.

Oganian and Karr (2006) suggest *microaggregation with adding noise*, which blends the clustering and perturbative effects of the two previous techniques. We set $\tilde{\boldsymbol{y}}_i^A = \tilde{\boldsymbol{y}}_{\text{mic},i}^A + \boldsymbol{\delta}_i$, where $\tilde{\boldsymbol{y}}_{\text{mic},i}^A$ is masked by microaggregation and $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^*)$. Oganian and Karr (2006) suggest using $\Sigma^* = \Sigma^A - \tilde{\Sigma}_{\text{mic}}^A$ (if this matrix is positive definite, and otherwise a positive definite approximation to it), where $\tilde{\Sigma}_{\text{mic}}^A$ denotes the sample covariance of $\{\tilde{\boldsymbol{y}}_{\text{mic},1}^A, \ldots, \tilde{\boldsymbol{y}}_{\text{mic},n}^A\}$. A variant of the method is using controlled noise with microaggregation (Shlomo and De Waal 2008): (i) define five subgroups by quintiles D_q where $q = 1, \ldots, 5$, (ii) partition records $i \in D_q$ into cluster $\mathcal{G}_{q,g}$ with size of τ_{cmic}, (iii) replace $\boldsymbol{y}_i^A$ with the group mean $\tilde{\boldsymbol{y}}_{\text{cmic},i}^A = \sum_{k \in \mathcal{G}_{q,g}} \boldsymbol{y}_k^A / \tau_{\text{cmic}}$, and (iv) produce final masked records by adding random noise, $\tilde{\boldsymbol{y}}_{\text{cmicn},i}^A = \tilde{\boldsymbol{y}}_{\text{cmic},i}^A + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^*)$ and Σ^* is the difference between the sample variance of $\{\boldsymbol{y}_j^A : j \in D_q\}$ and the sample variance of $\{\tilde{\boldsymbol{y}}_{\text{cmic},j}^A : j \in D_q\}$. We refer to this method as *controlled microaggregation with adding noise*. We note that the original paper of Shlomo and de Waal (2008) presents microaggregation for data with balance constraints; our version does not use the balance constraints.

Partially synthetic data (Rubin 1993; Little 1993; Reiter 2003) comprise the original n records with sensitive values replaced by multiple imputations. The imputations are generated from models estimated from the original data. The multiple copies enable data analyses to reflect imputation uncertainty appropriately. The additional data sets also offer more information for intruders to attempt identifications; see Reiter and Mitra (2009) and Drechsler and Reiter (2008) for further discussion of this issue.

2.2. *Approaches to SDL in the Presence of Edit Rules*

Both edit-after-SDL and edit-preserving SDL have potentially appealing features. Edit-after-SDL allows agencies to use existing SDL procedures and established edit-imputation procedures, including handling balance edits, without worrying about combining them. This may facilitate production operations when all edits are done in one step. On the other hand, edit-preserving SDL can reduce an agency's workload, since the masked data automatically satisfy the constraints. We now describe how one can implement these two strategies for the SDL methods outlined in Subsection 1. We note that, in some settings, it may be possible to use edit-preserving SDL for some constraints and edit-after-SDL for other constraints (e.g., Shlomo and De Waal 2008); we do not consider such mixed strategies here.

2.2.1. Approach I: Edit-After-SDL

In this approach, an agency first applies an SDL method to the collected data. Any post-SDL records that violate the constraints are deleted or "repaired" *ex post facto*. The agency treats any SDL-generated edit violations as if they were faulty values. This involves an error

localization step, for example, using the methods of Fellegi and Holt (1976), followed by replacing the localized errors with imputations that respect constraints. For example, one could use sequential regression imputation (Van Buuren and Oudshoorn 1999; Raghunathan et al. 2001), imputation from joint distributions (Geweke 1991; Tempelman 2007; Coutinho et al. 2011; Kim et al. 2014b), or in some settings hot-deck imputation (Bankier et al. 1994; Shlomo and De Waal 2005; Coutinho and De Waal 2012; Coutinho et al. 2013). As examples of this strategy, Shlomo and De Waal (2008) apply several SDL methods and correct edit-failing records via an edit-imputation procedure based on linear programming; and Cano and Torra (2011) propose adding random noise followed by swapping the noise values of edit-failing records until all records pass edit constraints. We note that neither of these approaches is theoretically guaranteed to preserve all edits.

To implement edit-after-SDL, we propose to use a model-based imputation method which guarantees that all edit-corrections result in records that lie in the feasible region, for example, the restricted support of $\boldsymbol{y}_i$ that satisfies all inequality constraints. Specifically, we adopt the multivariate imputation method proposed by Kim et al. (2014b), which is based on mixtures of multivariate normal distributions and is therefore flexible enough to describe complex distributional features. Let $\mathcal{Y}$ represent the feasible region. Using $K > 1$ mixture components – see Kim et al. (2014b) for discussion of setting K – we assume that

$$f(\boldsymbol{y}_i|\Theta_1, \ldots, \Theta_K) \propto \sum_{k=1}^{K} w_k \mathrm{N}(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \Omega_k) I(\boldsymbol{y}_i \in \mathcal{Y}). \quad (1)$$

Here, for each of the K mixture components, w_k is the probability (or weight) of the component, $(\boldsymbol{\mu}_k, \Omega_k)$ is the component mean vector and covariance matrix, and $\Theta_k = (w_k, \boldsymbol{\mu}_k, \Omega_k)$. After performing SDL, we identify each record with $\tilde{\boldsymbol{y}}_i \notin \mathcal{Y}$, blank its $\tilde{\boldsymbol{y}}_i^A$, and replace $\tilde{\boldsymbol{y}}_i^A$ with values generated from the posterior predictive distribution, $f(\boldsymbol{y}_i^A|D, \mathcal{Y})$. We refer readers to the Appendix for the specifications of the prior distributions and details of Markov chain Monte Carlo (MCMC) steps. We note that the imputation engine of Kim et al. (2014b) does not automatically extend to handle balance constraints, although it can be modified to do so (Kim et al. 2014a). We also note that agencies can ensure only integer values are released by rounding each imputed value to the nearest integer (we did not do this in our simulation).

2.2.2. Approach II: Edit-Preserving SDL

It is possible to modify some SDL techniques to ensure the masked data satisfy all constraints. A general strategy is to draw candidate masked values repeatedly until they satisfy all edit rules. This rejection sampling approach can be readily applied for SDL methods based on randomization, particularly when edit rules are based on sets of linear inequalities. For example, an agency that adds noise to variables can generate ε_i (or $\boldsymbol{\delta}_i$) repeatedly until the drawn $\tilde{\boldsymbol{y}}_i$ satisfies the edit rules. We note that rejection sampling approaches can have various negative impacts on data quality. For example, the distribution of the random noise for points near the boundary of the feasible region is not likely to be symmetric, which could result in bias. We also note that balance edits can be difficult to satisfy with rejection sampling.

For SDL methods not entailing randomization, rejection sampling is difficult to implement. Rejection sampling is not possible for typical implementations of microaggregation, since no randomization is involved in microaggregation, except possibly in clustering heuristics. Rejection sampling is generally inappropriate for partially synthetic data, since the model itself should account explicitly for the constrained support (the feasible region). Instead, we use the imputation engine of Kim et al. (2014b), heretofore used exclusively for missing data, as a synthesizer that guarantees the released synthetic values satisfy all edit constraints.

3. Simulation Study

We use a subset of 6,521 establishments from the 1991 Colombian Annual Manufacturing Survey data comprising seven numerical variables: number of skilled employees (SL), number of unskilled employees (UL), wages for skilled employees (SW), wages for unskilled employees (UW), value added (VA), material used in products (MU), and capital (CP). We assume that these records are error-free. As edit rules, we introduce linear constraints typical of those used to edit business survey data (Winkler and Draper 1996; Thompson et al. 2001; Hedlin 2003). Table 1 displays the range restrictions, and Table 2 displays the ratio constraints. The introduced constraints are data derived and hypothetical; they are not actual constraints derived from the domain knowledge of economic experts.

To simplify presentation, we mask only three of the seven variables – number of skilled employees, number of unskilled employees, and capital – and leave the remaining variables unaltered. We work with the natural logarithms of all variables. While not necessary, this improves computation in the mixture model used for imputations, as the model needs a smaller number of mixture components. Additionally, log transformations are often useful in statistical inference models with skewed economic data (Petrin and White 2011). To avoid new notation, we let $\boldsymbol{y}_i$ and $\tilde{\boldsymbol{y}}_i$ represent the vectors of natural logarithms of the seven variables in D and D^{rel}, respectively. Thus, $\boldsymbol{y}_i^A$ comprises the three log-transformed values $(y_{i\text{SL}}, y_{i\text{UL}}, y_{i\text{CP}})$.

We use the SDL procedures outlined in Section 2 on the log-transformed values $\boldsymbol{y}_i$, using multiple values of the disclosure parameters when possible. These include adding noise (Noise) with $\tau_{\text{noise}} \in \{0.16, 0.25, 0.36, 0.49\}$, rank swapping (Swap) with $\tau_{\text{swap}} \in \{1, 5, 10\}$, microaggregation based on principal components clustering (Mic)

Table 1. Description of variables in the 1991 Colombian Annual Manufacturing Survey with data-derived range restrictions

Variable	Label	Range restriction
Skilled labor	SL	0.9–400
Unskilled labor	UL	0.9–1,000
Wages paid to skilled labor	SW	300–3,000,000
Wages paid to unskilled labor	UW	600–4,000,000
Real value added	VA	50–1,000,000
Real material used in products	MU	10–1,000,000
Capital	CP	5–1,000,000

Table 2. Data-derived ratio edits ($V_1/V_2 \leq b$) for the 1991 Colombian Manufacturing Survey

	V_2						
V_1	SL	UL	SW	UW	VA	MU	CP
SL	1	20	0.01	0.01	0.1	0.3	2
UL	50	1	0.1	0.005	0.3	5	5
SW	20000	100000	1	50	300	500	1000
UW	66666.7	10000	100	1	200	5000	5000
VA	10000	20000	10	10	1	200	700
MU	50000	100000	33.3	100	100	1	1000
CP	20000	10000	10	16.7	100	100	1

with $\tau_{\text{mic}} \in \{2, 3, 5\}$, microaggregation based on principal components clustering followed by adding noise (MicN), and multivariate fixed-size microaggregation (MMic) with $\tau_{\text{mmic}} \in \{3, 10, 15, 30\}$. We also examined variable-size microaggregation (Solanas and Martnez-Balleste 2006; Domingo-Ferrer et al. 2008); the results were essentially indistinguishable from MMic with $\tau_{\text{mmic}} = 3$ and thus are not reported here. We also use two methods of Shlomo and de Waal (2008), including controlled adding noise (cNoise) with $\tau_{\text{cnoise}} \in \{0.10, 0.30, 0.50\}$ and controlled microaggregation with adding noise based on principal components clustering/subgrouping (cMicN) with $\tau_{\text{cmic}} \in \{2, 3, 5\}$. We generate partially synthetic data (Synt) by replacing all of $\boldsymbol{y}_i^A$ with draws from the model of Kim et al. (2014b). For partially synthetic data, we use only a single draw of the parameters from a converged Markov chain to generate one realization of D^{rel}; in practice, we recommend using multiple draws and releasing multiple data sets to enable variance estimation, provided that doing so does not increase risks unacceptably.

For procedures involving randomness, we generate 20 masked data sets from different random seeds. For the microaggregation procedures (Mic and MMic), we use only one masked data set since these methods are deterministic. As evident in Table 3 and illustrated in Figure 1, all the perturbative SDL methods except MMic3 and MMic10 result in edit violations when applied without edit-preserving modifications. Adding noise with the larger values of τ_{noise} pushes many $\boldsymbol{y}_i$ outside the boundary of $\mathcal{Y}$, resulting in the largest number of edit violations. Rank swapping also produces many edit violations, even with the fairly tight swapping range of $\tau_{\text{swap}} = 10$. Microaggregation and multivariate

Table 3. Numbers of records that violate edit rules across the 20 replications (or single realizations for Mic and MMic) after implementing perturbative SDL methods

Method	Mean	%	%	Mean	%	Method	Mean	%
Noise16	157.8	2.5	Mic3N	84.1	1.3	Mic2	4.0	0.1
Noise25	255.4	4.0	Mic5N	116.2	1.8	Mic3	5.0	0.1
Noise36	406.2	6.3	cMic2N	54.8	0.8	Mic5	15.0	0.2
Noise49	614.8	9.6	cMic3N	83.1	1.2	MMic3	0.0	0.0
cNoise10	7.6	0.1	cMic5N	116.1	1.8	MMic10	0.0	0.0
cNoise30	27.9	0.4	Swap01	5.6	0.1	MMic15	1.0	0.02
cNoise50	48.1	0.7	Swap05	45.1	0.7	MMic30	2.0	0.03
Mic2N	53.5	0.8	Swap10	134.2	2.1			

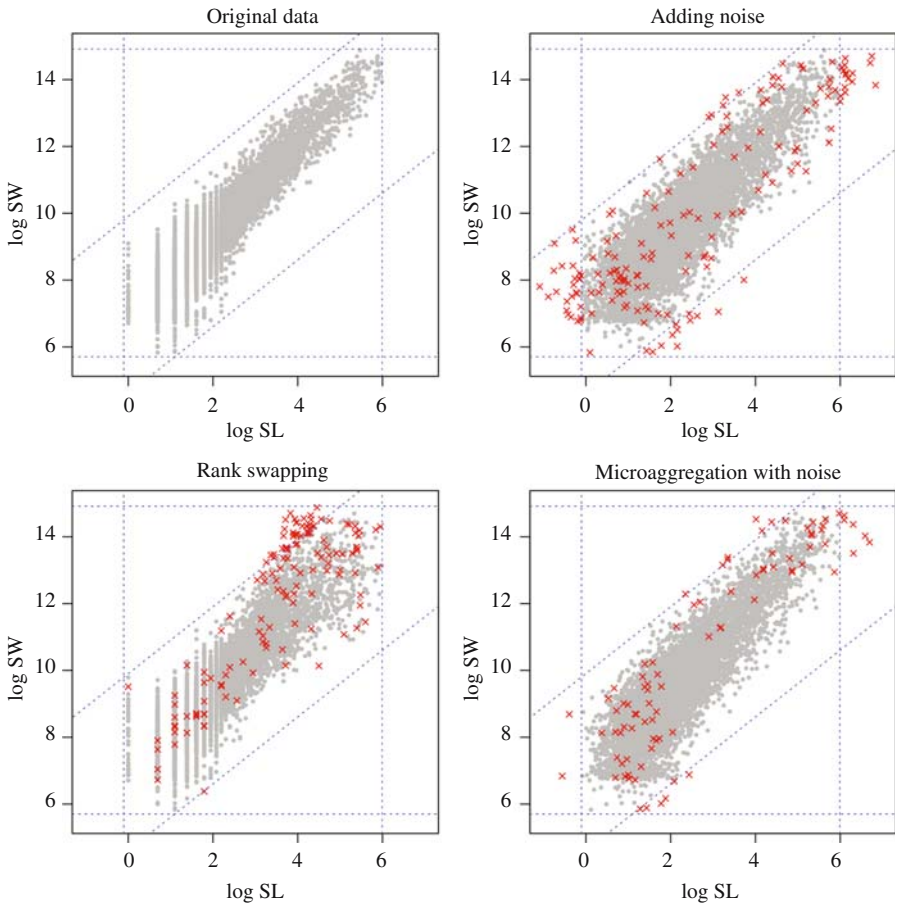

Fig. 1. Illustrative example of how SDL can result in violations of linear constraints. Top-left panel shows pre-SDL data for the log(SL) and log(SW). The variables SL, UL, and CP are masked by adding noise with $\tau_{noise} = 0.16$ (Noise16, top-right panel), rank swapping with $\tau_{swap} = 10$ (Swap10, bottom-left panel), and microaggregation of $\tau_{mic} = 3$ with adding noise (Mic3N, bottom-right panel). Solid circles indicate records that satisfy edit rules and " × " indicate records that violate constraints, i.e., $\tilde{\mathbf{y}}_i \notin Y$

fixed-size microaggregation result in only a few masked records that violate the constraints. This is because microaggregation generally moves values away from boundaries and hence towards the feasible region. In fact, if we had applied microaggregation to all variables in $\mathbf{y}_i$, the resulting records always would be inside $\mathcal{Y}$ due to its convexity. Since we replace only each $\mathbf{y}_i^A$, we cannot guarantee that $\tilde{\mathbf{y}}_i \in Y$. As a general conclusion, we note that the number of edit violations increases with the amount of perturbation for every class of SDL methods.

We next seek to correct any edit violations using the two general strategies. For edit-after-SDL, we replace all values of $\mathbf{y}_i^A$ of edit-failing records with draws from the imputation model outlined in Subsection 2.2.1. For edit-preserving SDL, we use the rejection sampling scheme of Subsection 2.2.2 for all methods involving randomness. For rank swapping with $\tau_{\text{swap}} = 10$, we did not obtain a D^{rel} without edit violations even after 1,000 independent replications of swapping. Each D^{rel} had at least 99 out of 6,521 records

that violated the constraints, suggesting that waiting for a constraint-preserving, rank-swapped data set for this procedure in this simulation design is hopeless.

As measures of disclosure risk, we use the *percentage of linked* criterion of Domingo-Ferrer, Mateo-Sanz, and Torra (2001). First, we compute the distances

$$d_{i,j} = \sqrt{\sum_{l} (y_{il}^{A} - \tilde{y}_{jl}^{A})^2}, \qquad \forall i,j = 1, \ldots, n,$$

where $l \in$ (SL, UL, CP). For each i, we find the record j that achieves the minimum value of $d_{i,j}$. When $y_{i0} = y_{j0}$, that is, the record in D^{rel} can be linked correctly to D based on matching the available variables, we let $t_i^{(1)} = 1$ and otherwise let $t_i^{(1)} = 0$. We then define one risk measure as PL1 $= \sum_{i=1}^{n} t_i^{(1)}/n \times 100$. Similarly, we let $t_i^{(2)} = 1$ when the correct link for record i in D has either the smallest or second smallest value among all the $d_{i,j}$, and $t_i^{(2)} = 0$ otherwise. We define a second risk measure as PL2 $= \sum_{i=1}^{n} t_i^{(2)}/n \times 100$, the percentage of records for which the correct link is among the two closest matches. Finally, we define a third risk measure, PL3, as the percentage of records for which the correct link is among the three closest matches.

We use two measures of data utility: an approximate Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) of D^{rel} from D, and the propensity score (U_{prop}) utility measure suggested by Woo et al. (2009). For KL, we use a closed-form expression based on a normality assumption,

$$\text{KL} = \frac{1}{2}\left[\text{tr}\left\{(\Sigma^{\text{rel}})^{-1}\Sigma\right\} + \left(\bar{y}^{\text{rel}} - \bar{y}\right)^{T}(\Sigma^{\text{rel}})^{-1}\left(\bar{y}^{\text{rel}} - \bar{y}\right) - p - \log\left(\frac{|\Sigma^{\text{rel}}|}{|\Sigma|}\right)\right], \quad (2)$$

where $\bar{y}$ and Σ are the sample mean and the sample covariance of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ in D, and $\bar{\boldsymbol{y}}^{\text{rel}}$ and Σ^{rel} are the corresponding statistics of $\{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n\}$ in D^{rel}. For U_{prop}, we first concatenate D^{rel} and D, and add an indicator variable whose values equal one for all records in D^{rel} and equal zero for all records in D. Using the concatenated data, we estimate the logistic regression of the indicator variable on all seven variables (after log transformations), including main effects and all interactions up to third order; that is, we fit

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{a=1}^{7} \beta_a \log Y_{ia} + \sum_{a,b} \beta_{ab} \log Y_{ia} \log Y_{ib} + \sum_{a,b,c} \beta_{abc} \log Y_{ia} \log Y_{ib} \log Y_{ic}.$$

For $i = 1, \ldots, 2n$, we compute the set of predicted probabilities $\hat{p}_i$. The utility measure is

$$U_{\text{prop}} = \frac{1}{2n} \sum_{i=1}^{2n} \left(\hat{p}_i - \frac{1}{2}\right)^2.$$

Values of U_{prop} near zero represent high data utility, since they imply we are not able to distinguish between D^{rel} and D.

Table 4 displays the average values of KL, U_{prop} and PL1 — PL3 over the replicates for each method. When methods are implemented with both strategies, the risk-utility profiles are fairly similar across the two strategies. This is not overly surprising, since these SDL methods typically generate only a modest number of edit violations in these data. Nonetheless, for these methods, the edit-after-SDL version does slightly outperform the edit-preserving SDL version, generally offering both lower risk and higher utility. This results largely from the imputations, which are generally of higher quality than the repeated draws from the rejection sampling scheme.

In Table 4, the differences in the risk-utility profiles across the two ways of dealing with edit violations are dwarfed by differences in the profiles across the classes of SDL methods. This suggests that the choice of SDL method is more important than the strategy for correcting edit violations.

Figure 2 displays a risk-utility (R-U) map (Duncan and Stokes 2004; Gomatam et al. 2005; Cox et al. 2011) for all realizations of D^{rel} and the most competitive procedures, using U_{prop} as the utility measure and PL1 as the risk measure. The risk-utility frontier consists of candidate releases with no other candidate to their "southwest." The R-U frontier includes the variants of microaggregation with adding noise (MicN), which have the lowest levels of disclosure risk, and partially synthetic data (Synt), which has the maximum level of data utility and a low level of disclosure risk. Several variants of MMic are close to the frontier (and would be on the frontier but for Synt and Swap10), generally having high utility for reasonable disclosure risks.

4. Concluding Remarks

Based on our studies, there appear to be no appreciable differences between the strategies of edit-after-SDL and edit-preserving SDL, at least when both are possible. Hence, arguably, agencies can choose an SDL procedure without too much consideration of how they will ensure the released data satisfy all edits, at least when the SDL method does not generate a large number of edit violations. Microaggregation with adding noise, multivariate fixed-size microaggregation and partially synthetic data were the most effective strategies in our simulations. The last method has the additional advantage that the synthesis methodology can be used to impute missing data values and implement edit-preserving SDL simultaneously, following the two-stage approach described in Reiter (2004).

An intriguing aspect of the editing–SDL "disconnect" is whether edited values should be protected in the same way as original reported data. This point, perhaps, is more subtle than it may seem initially. One interpretation is that a statistical agency promises to protect whatever information the subjects provide, even if that information is believed, or known to be, erroneous. Under this logic, edited and imputed values are not respondent information (i.e., they have been imputed rather than reported) and therefore might be treated differently during SDL. Another view is that the agency is also charged with protecting its best estimate of actual values, as opposed to reported values, which implies that edited and imputed values do require SDL. To our knowledge this issue remains unresolved and, indeed, largely unaddressed. We believe that in the long run, the most desirable approach is one that fully integrates editing, imputation and SDL.

Table 4. Measured data utility and disclosure risk. Entries include the averages of KL, U_{prop}, PL1, PL2, and PL3 from 20 replications of each method, except Mic and MMic which have only single replicates. Note that the three risk measures are highly correlated (all correlations are at least 0.988), so that they offer similar conclusions about the risk-utility profiles of the different SDL methods

	Inverse Utility				Risk					
	KL		U_{prop} (×100)		PL1		PL2		PL3	
Methods	I	II	I	II	I	II	I	II	I	II
Noise16	0.34	0.35	2.2	2.2	1.91	2.12	3.42	3.75	4.74	5.20
Noise25	0.50	0.52	2.9	3.0	1.11	1.26	2.05	2.26	2.90	3.20
Noise36	0.64	0.67	3.4	3.5	0.74	0.82	1.37	1.52	1.94	2.12
Noise49	0.75	0.81	3.5	3.8	0.51	0.60	0.96	1.10	1.38	1.55
cNoise10	0.0007	0.0007	0.002	0.001	48.21	48.34	67.32	67.45	77.47	77.59
cNoise30	0.04	0.04	0.05	0.05	8.78	8.85	14.91	15.07	19.83	20.00
cNoise50	0.16	0.16	0.2	0.3	2.56	2.60	4.63	4.65	6.47	6.45
Mic2N	0.50	0.51	2.9	3.0	0.57	0.58	1.10	1.12	1.61	1.64
Mic3N	0.64	0.66	4.0	4.2	0.35	0.35	0.67	0.66	0.99	0.97
Mic5N	0.75	0.78	4.8	5.2	0.29	0.24	0.54	0.45	0.76	0.66
cMic2N	0.51	0.51	2.9	3.0	0.57	0.59	1.10	1.11	1.61	1.64
cMic3N	0.64	0.66	4.1	4.3	0.34	0.32	0.69	0.65	0.99	0.97
cMic5N	0.75	0.79	4.9	5.3	0.27	0.24	0.52	0.45	0.78	0.65
Swap01	0.002	0.002	0.004	0.004	63.55	63.81	79.79	80.22	85.91	86.27
Swap05	0.08	–	0.05	–	6.22	–	12.01	–	17.26	–
Swap10	0.24	–	0.1	–	0.94	–	1.86	–	2.91	–
Mic2	0.59	–	2.7	–	1.38	–	3.05	–	4.18	–
Mic3	1.34	–	4.6	–	0.67	–	1.43	–	2.19	–
Mic5	2.71	–	6.3	–	0.34	–	0.75	–	1.24	–
MMic3	–	0.01	–	0.01	–	7.73	–	15.83	–	26.83
MMic10	–	0.05	–	0.1	–	2.12	–	4.22	–	6.17
MMic15	0.08	–	0.2	–	1.37	–	2.75	–	4.28	–
MMic30	0.16	–	0.3	–	0.74	–	1.56	–	2.37	–
Synt	–	0.02	–	0.06	–	0.59	–	1.13	–	1.66

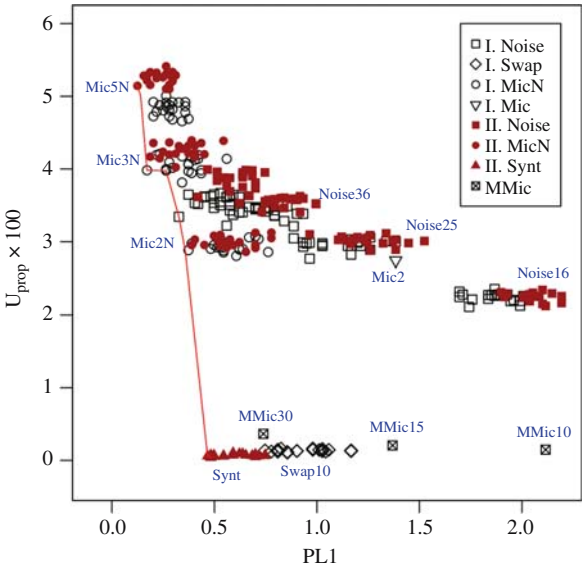

Fig. 2. Risk-utility map with the SDL methods. The solid line indicates the risk-utility frontier. The open symbols represent edit-after-SDL approaches, and the solid symbols represent edit-preserving SDL approaches. Smaller values of PL1 and U_{prop} represent the higher levels of data protection and data utility. Note that the plot does not include cMicN's because the results are very similar to those of MicN. The other methods whose results are not shown in the plot have high risk and/or low utility

Finally, we note two somewhat technical issues. First, some statistical agencies do not always include edit and imputation flags in released data. The risk and utility consequences of doing this are unexplored. The underlying issue is one of transparency (Karr 2009; Cox et al. 2011). Second, our research to date has not touched the role of weights, which was addressed to some extent in Cox et al. (2011). Weights themselves may pose disclosure risk (e.g., of unreleased values of design variables), but are generally ignored in all three of the editing, imputation and SDL processes. Some editing procedures, such as seeking additional information from "large" and low–weight respondents, consider weights implicitly. Some implementations of data swapping can accommodate weight constraints. Indexed microaggregation (Cox et al. 2011) is able to protect risky weights. However, by any measure, much more work remains than has been carried out so far.

Appendix: The Joint Multivariate Imputation Using Normal Mixtures

For imputations of faulty values, we use the joint multivariate normal method developed in Kim et al. (2014b) and described in Section 2. The likelihood function in (1) can be re-expressed with latent variables z_i by

$$f(\mathbf{y}_i|z_i, \mu, \Omega) \propto \mathrm{N}(\mathbf{y}_i|\boldsymbol{\mu}_{z_i}, \Omega_{z_i})I(\mathbf{y}_i \in \mathcal{Y})$$

and

$$Pr(z_i = k) = w_k, k = 1, \ldots, K.$$

Following Lavine and West (1992), we assume the prior distributions,

$$\boldsymbol{\mu}_k \,|\, \Omega_k \sim \mathrm{N}(\boldsymbol{\mu}_0, h^{-1}\Omega_k), \quad \Omega_k \sim \mathrm{IW}(\zeta, \Phi)$$

where $\Phi = diag(\phi_1, \ldots, \phi_p)$, and $\phi_j \sim \mathrm{Gamma}(a_\phi, b_\phi)$ for $j = 1, \ldots, p$. Here, IW denotes the inverse Wishart distribution and Gamma(a,b) denotes the Gamma distribution with mean a/b. For flexible modeling of the component weights, we adopt the stick-breaking representation of a truncated Dirichlet process (Sethuraman 1994; Ishwaran and James 2001):

$$\begin{aligned} w_k &= v_k \prod_{g<k} (1 - v_g) \text{ for } k = 1, \ldots, K \\ v_k &\sim \mathrm{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1;\ v_K = 1 \\ \alpha &\sim \mathrm{Gamma}(a_\alpha, b_\alpha). \end{aligned}$$

In the simulation study, we follow Kim et al. (2014b) and set $\boldsymbol{\mu}_0 = 0$, $h = 1$, $\zeta = p + 1$, $a_\phi = b_\phi = 0.25$, $a_\alpha = b_\alpha = 0.25$ and $K = 40$.

To facilitate the estimation of μ and Ω, we use a data-augmentation technique developed by O'Malley and Zaslavsky (2008). The data augmentation supposes a larger, hypothetical sample $Y_N = \{Y_n, Y_{N-n}\}$ where Y_n is the set of $\boldsymbol{y}_i \in \mathcal{Y}$ following the likelihood in Equation (1) and Y_{N-n} consists of the values from outside of $\mathcal{Y}$, so that

$$f(Y_N \,|\, \Theta_1, \ldots, \Theta_K) = \prod_{i=1}^{N} \sum_{k=1}^{K} w_k \mathrm{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \Omega_k),$$

where $\Theta_k = (\mu_k, \Omega_k, w_k)$. Given the augmented sample Y_N, the parameters $\Theta_k = (w_k, \mu_k, \Omega_k)$ can be sampled via Gibbs sampling. Setting $f(N) \propto 1/N$ as suggested by Meng and Zaslavsky (2002) and O'Malley and Zaslavsky (2008), the conditional density of the size of Y_{N-n} is distributed as

$$N - n \,|\, n, \Theta_1, \ldots, \Theta_K, \mathcal{Y} \sim \text{Negative Binomial}\big(n, 1 - h_\Theta(\mathcal{Y})\big),$$

where

$$h_\Theta(\mathcal{Y}) = \int_{\{\boldsymbol{y}: \boldsymbol{y} \in \mathcal{Y}\}} \sum_{k=1}^{K} w_k \mathrm{N}(\boldsymbol{y} | \boldsymbol{\mu}_k, \Omega_k) d\boldsymbol{y}.$$

The MCMC algorithm for sampling from this distribution relies on the following steps.

1. For $k = 1, \ldots, K$, draw $\Omega_k \sim \mathrm{IW}(\zeta_k, \Phi_k)$ and $\boldsymbol{\mu}_k \sim \mathrm{N}(\boldsymbol{\mu}_k^*, \Omega_k/(N_k + h))$ where $\boldsymbol{\mu}_k^* = (N_k \bar{\boldsymbol{y}}_k + h\boldsymbol{\mu}_0)/(N_k + h)$, $\zeta_k = \zeta + N_k$,$\Phi_k = \Phi + S_k + (\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)'/(1/N_k + 1/h)$. We calculate the sample mean $\bar{\boldsymbol{y}}_k$ and the sample covariance S_k from the error-free, pre-SDL values $Y_n = \{\boldsymbol{y}_i, \mathrm{i} = 1, \ldots, n\}$ and the drawn auxiliary values Y_{N-n} by $\bar{\boldsymbol{y}}_k = \sum_{\{i; z_i = k\}} \boldsymbol{y}_i / N_k$ where $N_k = \sum_{i=1}^{N} I(z_i = k)$ and $S_k = \sum_{\{i; z_i = k\}} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)'$.
2. For $k = 1, \ldots, K-1$, draw $v_k \sim \mathrm{Beta}\Big(1 + N_k, \alpha + \sum_{g>k} N_g\Big)$. Set $v_K = 1$. Compute $w_k = v_k \prod_{g<k} (1 - v_g)$.

3. Update $\Phi = diag(\phi_1, \ldots, \phi_p)$ by drawing $\phi_j \sim \text{Gamma}\left(a_\phi + \zeta K/2, b_\phi + \sum_{k=1}^{K} \Omega_{k(j,j)}^{-1}/2\right)$ for each $j = 1, \ldots, p$, where $\Omega_{k(j,j)}^{-1}$ is the jth diagonal element of Ω_k^{-1}.
4. Draw α from $\text{Gamma}\left(a_\alpha + K - 1, b_\alpha - \log w_K\right)$.
5. For $i = 1, \ldots, n$, sample $z_i \sim \text{Categorical}(w_{i1}^*, \ldots, w_{iK}^*)$ where

$$w_{ik}^* = w_k \text{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \Omega_k) / \left[\sum_{g=1}^{K} w_g \text{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_g, \Omega_g)\right].$$

6. Sample (N, Z_{N-n}, Y_{N-n}) jointly from their full conditional distribution as follows. Let $c_{\text{in}} = c_{\text{out}} = 0$.
   6.1. Draw $z^* \sim \text{Categorical}(w_1, \ldots, w_K)$.
   6.2. Draw $\boldsymbol{y}^* \sim \text{N}(\boldsymbol{\mu}_z^*, \Omega_{z^*})$.
   6.3. If $\boldsymbol{y}^* \in \mathcal{Y}$, set $c_{\text{in}} = c_{\text{in}} + 1$.
   6.4. If $\boldsymbol{y}^* \in \mathcal{Y}^c$, set $c_{\text{out}} = c_{\text{out}} + 1$, $\boldsymbol{y}_{n + c_{\text{out}}} = \boldsymbol{y}^*$, and $z_{n+c_{\text{out}}} = z^*$.
   6.5. Repeat 6.1 through 6.3 until $c_{\text{in}=n}$.

   Let $N = n + c_{\text{out}}$. Now, $Y_{N-n} = \{\boldsymbol{y}_n + 1, \ldots, \boldsymbol{y}_n + c_{\text{out}}\}$ and $Z_{N-n} = \{z_{n+1}, \ldots, z_{n+c_{\text{out}}}\}$.
7. To update the replacement draws of the faulty values, we use a Hit-and-Run sampler (Chen and Schmeiser 1993). In the initialization step, we propose a starting value $\tilde{\boldsymbol{y}}_i^{A(0)}$ such that $(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(0)}) \in \mathcal{Y}$, for example by using rejection sampling or an extreme-points approach (see Kim et al. 2014b). At any MCMC iteration $t \geq 0$, we update the current value $\tilde{\boldsymbol{y}}_i^{A(t)}$ (which replaces the faulty $\tilde{\text{y}}_i^A$) with the following steps.
   7.1. Draw a direction $\boldsymbol{d}^*$ uniformly from the surface of the $|\tilde{y}_i^A|$-dimensional unit sphere centered at the origin.
   7.2. Draw a signed distance λ^* from the uniform distribution on Ξ,

$$\Xi = \left\{\lambda : \left(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(t)} + \lambda \boldsymbol{d}^*\right) \in \mathcal{Y}\right\}$$

   7.3. Accept or reject the proposal $\tilde{\boldsymbol{y}}_i^{A*} = \tilde{\boldsymbol{y}}_i^{A(t)} + \lambda^* \boldsymbol{d}^*$ with the acceptance probability ρ_i, where

$$\rho_i = \min\left[1, \frac{f(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A*} | \Theta_{z_i})}{f(\boldsymbol{y}_i^U, \tilde{\boldsymbol{y}}_i^{A(t)} | \Theta_{z_i})}\right].$$

5. References

Bankier, M., M. Luc, C. Nadeau, and P. Newcombe. 1994. "Imputing Numeric and Qualitative Variables Simultaneously." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 242–247. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1994_036.pdf. (accessed February 2015).

Cano, I. and V. Torra. 2011. "Edit Constraints on Microaggregation and Additive Noise." In *Privacy and Security Issues in Data Mining and Machine Learning*, edited by

C. Dimitrakakis, A. Gkoulalas-Divanis, A. Mitrokotsa, V.S. Verykios, and Y. Saygin, 1–14. Berlin: Springer.

Chen, M.H. and B. Schmeiser. 1993. "Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers." *Journal of Computational and Graphical Statistics* 2: 251–272. DOI: http://dx.doi. org/10.2307/1390645.

Coutinho, W. and T. de Waal. 2012. *Hot Deck Imputation of Numerical Data Under Edit Restrictions*. Discussion Paper 2012243, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/6C97F296-EE33-4F26-A813-6432ED530249/0/201223x10pub.pdf. (accessed February 2015).

Coutinho, W., T. de Waal, and M. Remmerswaal. 2011. "Imputation of Numerical Data Under Linear Edit Restrictions." *Statistics and Operations Research Transactions* 35: 29–62.

Coutinho, W., T. de Waal, and N. Shlomo. 2013. "Calibrated Hot-Deck Donor Imputation Subject to Edit Restrictions." *Journal of Official Statistics* 29: 299–321. DOI: http://dx.doi.org/10.2478/jos-2013-0024.

Cox, L.H., A.F. Karr, and S.K. Kinney. 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.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: Wiley.

Defays, D. and P. Nanopoulos. 1993. "Panels of Enterprises and Confidentiality: The Small Aggregates Method." In Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, November 2–4, 1992, 195–204. Ottawa, Ontario, Canada. Available at: http://www.researchgate. net/publication/243784453_Panels_of_enterprises_and_confidentiality_the_small_aggregates_ method. (accessed February 2015).

Domingo-Ferrer, J. and J.M. Mateo-Sanz. 2002. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering* 14: 189–201. DOI: http://dx.doi.org/10.1109/69.979982.

Domingo-Ferrer, J., F. Sebe, and A. Solanas. 2008. "A Polynomial-Time Approximation to Optimal Multivariate Microaggregation." *Computers and Mathematics with Applications* 55: 714–732. DOI: http://dx.doi.org/10.1016/j.camwa.2007.04.034.

Domingo-Ferrer, J., J.M. Mateo-Sanz, and V. Torra. 2001. "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk." In Pre-proceedings of ENKNTTS, 807–826. Available at: http://neon.vb.cbs.nl/casc/NTTSJosep.pdf. (accessed February 2015)

Drechsler, J. and J.P. Reiter. 2008. "Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and Y. Saygin, 227–238. New York: Springer.

Duncan, G.T. and S.L. Stokes. 2004. "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding." *Chance* 17: 16–20. DOI: http://dx.doi.org/10.1080/09332480.2004.10554908.

Fayyoumi, E. and B.J. Oommen. 2010. "A Survey on Statistical Disclosure Control and Microaggregation Techniques for Secure Statistical Databases." *Software: Practice and Experience* 40: 1161–1188. DOI: http://dx.doi.org/10.1002/spe.992.

Fellegi, I.P. and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association* 71: 17–35. DOI: http://dx.doi.org/10.1080/01621459.1976.10481472.

Geweke, J. 1991. "Efficient Simulation from the Multivariate Normal and Student-T Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities." In Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, April 21–24, 1991. 571–578. Seattle, Washington. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.27.568&rep = rep1&type = pdf. (accessed February 2015).

Gomatam, S., A.F. Karr, J.P. Reiter, and A.P. Sanil. 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.

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

Hedlin, D. 2003. "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics." *Journal of Official Statistics* 19: 177–199.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control*. West Sussex, UK: John Wiley & Sons.

Ishwaran, H. and L.F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96: 161–173. DOI: http://dx.doi.org/10.1198/016214501750332758.

Karr, A.F. 2009. *The Role of Transparency in Statistical Disclosure Limitation*. Presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.41.e.pdf. (accessed February 2015).

Kim, H.J., L.H. Cox, A.F. Karr, J.P. Reiter and Q. Wang. 2014a. *Simultaneous Edit-Imputation for Contineous Microdata*. Technical Report 189, National Institute of Statistical Sciences, Research Triangle Park, NC. Available at: https://www.niss.org/sites/default/files/tr189_updated.pdf (accessed February 2015).

Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox, and A.F. Karr. 2014b. "Multiple Imputation of Missing or Faulty Values Under Linear Constraints." *Journal of Business & Economic Statistics* 32: 375–386. DOI: http://dx.doi.org/10.1080/07350015.2014.885435.

Kim, J.J. 1986. "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 370–374. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1986_069.pdf. (accessed February 2015).

Kullback, S. and R.A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22: 79–86.

Lavine, M. and M. West. 1992. "A Bayesian Method for Classification and Discrimination." *Canadian Journal of Statistics* 20: 451–461. DOI: http://dx.doi.org/10.2307/3315614.

Little, R.J.A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9: 407–426.

Meng, X.L. and A.M. Zaslavsky. 2002. "Single Observation Unbiased Priors." *The Annals of Statistics* 30: 1345–1375.

Moore, R.A. 1996. *Controlled Data-Swapping Techniques for Masking Use Microdata Sets*. Research Report RR96/04, Statistical Research Division, U.S. Bureau of the Census, Washington, DC. Available at: https://www.census.gov/srd/papers/pdf/rr96-4.pdf. (accessed February 2015).

Oganian, A. and A.F. Karr. 2006. "Combinations of SDC Methods for Microdata Protection." In *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer and L. Franconi. 102–113. Berlin: Springer.

O'Malley, A.J. and A.M. Zaslavsky. 2008. "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse." *Journal of the American Statistical Association* 103: 1405–1418. DOI: http://dx.doi.org/10.1198/016214508000000724.

Petrin, A. and T.K. White. 2011. "The Impact of Plant-Level Resource Reallocations and Technical Progress on U.S. Macroeconomic Growth." *Review of Economic Dynamics* 14: 3–26. DOI: http://dx.doi.org/10.1016/j.red.2010.09.004.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27: 85–95.

Reiter, J.P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29: 181–188.

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

Reiter, J.P. 2005. "Estimating Risks of Identification Disclosure in Microdata." *Journal of the American Statistical Association* 100: 1103–1112. DOI: http://dx.doi.org/10.1198/016214505000000619.

Reiter, J.P. and R. Mitra. 2009. "Estimating Risks of Identification Disclosure in Partially Synthetic Data." *Journal of Privacy and Confidentiality* 1: 99–110.

Rubin, D.B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9: 461–468.

Sethuraman, J. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4: 639–650.

Shlomo, N. and T. de Waal. 2005. *Preserving Edits When Perturbing Microdata for Statistical Disclosure Control*. S3RI Methodology Working Paper M05/12, Southampton Statistical Sciences Research Institute. Available at: http://eprints.soton.ac.uk/14725/1/14725-01.pdf. (accessed February 2015).

Shlomo, N. and T. de Waal. 2008. "Protection of Micro-Data Subject to Edit Constraints Against Statistical Disclosure." *Journal of Official Statistics* 24: 229–253.

Solanas, A. and A. Martnez-Balleste. 2006. "V-MDAV: A Multivariate Microaggregation With Variable Group Size." In Proceedings of the 17th IASC Symposium on Computational Statistics, August 28–September 1, 2006. 917–925. Rome, Italy. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.1680&rep=rep1&type=pdf. (accessed February 2015).

Sullivan, G. and W.A. Fuller. 1989. "The Use of Measurement Error to Avoid Disclosure." In Proceedings of the Section on Survey Research Method of the American Statistical Association, 802–807. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1989_148.pdf. (accessed February 2015).

Tempelman, C. 2007. *Imputation of Restricted Data*. Ph. D. dissertation, University of Groningen. Available at: http://dissertations.ub.rug.nl/faculties/eco/2007/d.c.g.tempelman. (accessed February 2015).

Tendick, P. 1991. "Optimal Noise Addition for Preserving Confidentiality in Multivariate Data." *Journal of Statistical Planning and Inference* 27: 341–353. DOI: http://dx.doi.org/10.1016/0378-3758(91)90047-I.

Thompson, K.J., K. Sausman, M. Walkup, S. Dahl, C. King, and S.A. Adeshiyan. 2001. *Developing Ratio Edits and Imputation Parameters for the Services Sector Censuses Plain Vanilla Ratio Edit Module Test*. Economic Statistical Methods Report ESM-0101, U.S. Bureau of the Census, Washington, DC.

Torra, V. 2008. "Constrained Microaggregation." *Transactions on Data Privacy* 1: 86–104.

Van Buuren, S. and K. Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden, Netherlands. Available at: http://www. stefvanbuuren.nl/publications/Flexible%20multivariate%20-%20TNO99054%201999.pdf. (accessed February 2015)

Willenborg, L. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Winkler, W.E. and L.R. Draper. 1996. *Application of the SPEER Edit System*. Research Report RR96/02, Statistical Research Division, U.S. Bureau of the Census, Washington, DC. Available at: https://www.census.gov/srd/papers/pdf/rr96-2.pdf. (accessed February 2015).

Woo, M.J., J.P. Reiter, A. Oganian, and A.F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1: 111–124.

Received October 2013
Revised May 2014
Accepted September 2014

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 139–140, http://dx.doi.org/10.1515/JOS-2015-0007

Book Review

Morgan S. Earp[1]

Cristina Davino and Luigi Fabbris. *Survey Data Collection and Integration*. 2013 Berlin: Springer-Verlag, ISBN 978-3-642-21307-6, 155 pp, $109.

As editors of the book *Survey Data Collection and Integration,* Davino and Fabbris provide a collection of papers presenting practical solutions to real problems in statistical surveys. The papers included in the book discuss survey challenges such as questionnaire design, record linkage, imputation, and calibration weighting. The papers contained in this text proceeded from discussions arising during the "Thinking about Methodology and Applications of Surveys Workshop" at the University of Macerata (Italy) in September of 2010. With only 155 pages, the book reads like a special conference issue of *JOS.* All of the papers provide a review of the related literature, highlight a statistical survey challenge, and describe a case-study solution that can be applied by practitioners and studied by academics.

In Part One of the book, Biggeri provides an introduction to statistical surveys and discusses two different frameworks used to assess the quality of statistical surveys: 1) the total quality management approach (Groves 1989; Groves and Tortora 1991); and 2) the life cycles of surveys from a quality perspective (Groves et al. 2009). Biggeri highlights critical issues, challenges, and the need for development in statistical surveys; specifically focusing on mode of data collection, questionnaire construct, sample design, estimation, respondent burden, data discrimination, and standardization. Biggeri stresses the importance of uniting the efforts of both practitioners and academics in order to not find only the optimal, but also the most practical solutions to the challenges faced by statistical surveys. The remainder of the book is authored by both university and government researchers, thus providing both the academic and practitioner perspective on survey and measurement challenges – integrating both theory and real-world solutions.

Part Two of the book highlights tools used by psychometricians to evaluate questionnaire design. Fabbris discusses how to rank items, pick the best/worst items, and compare items based on the survey procedures, the type of scale being used, respondent burden, missing data, and data collection mode. Davino and Romano provide an innovative approach for assessing multi-item subjective measurement scales. As opposed to taking a more advanced psychometric approach to assess differences among items such as structural-equation modeling or item-response theory, the authors aim to assess different subjective-scale measurement items using mixed-model ANOVA (McCulloch and Searle 2001) and multivariate methods (Mardia et al. 1979), which allow for the

[1] U.S. Bureau of Labor Statistics, Office of Survey Methods Research, 2 Massachusetts Avenue, NE 1950, Washington, DC 20212, U.S.A. Email: Earp.Morgan@bls.gov

comparison of different multi-item scales while considering the information provided by each single item within a scale. While this is possible using item response theory, these methods are more familiar and easily understood by survey practitioners with little to no background in psychometric theory. Balbi and Triunfo describe statistical tools used to jointly analyze closed and open-ended questions. Lastly, Napoli and Arcidiancono explore the use of self-anchoring scales in social research in terms of measuring attitudes and opinions and constructing a self-anchoring scale; ultimately this paper provides a case study highlighting the utility and applicability of self-anchoring questions in survey research that allows the participants' opinions of their abilities to prevail over that of the researchers'. This part of the book provides a light overview of psychometric theory and demonstrates how its concepts can be used to evaluate and compare statistical survey items.

Part Three and Four of the book focus on data integration and weighting to adjust for missing data and nonresponse. Part Three discusses sampling design and error estimation in relation to small-area estimation of poverty indicators (Pratesi, Giusti, and Marchetti), nonsampling errors in household surveys (D'Alessio and Ilardi), and the process of enriching large scale surveys through data fusion (Aluja-Banet, Daunis-i-Estadella, and Chen). In Part Four, Bellisai, Fivizzani, and Sorrentino explore different methods used to integrate data across multiple business surveys in order to eliminate missing data and the use of calibration weighting to adjust for nonresponse bias after imputation is complete.

This book provides an interesting set of case studies that have integrated the work of both academics and practitioners to address the prevalent statistical survey challenges faced by survey methodologists. This book is a recommended read for practitioners interested in making use of the item assessment tools developed by psychometricians and those interested in using record linkage across multiple surveys to reduce item missingness. Just like a special issue of JOS, this book's strength lies in the integration of theory and case studies highlighting real-world specific problems currently faced by survey practitioners. Since each paper is so specifically focused, it would be recommended as a text for advanced students and/or current survey practitioners.

References

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

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

Groves, R.M., and R.D. Tortora. 1991. Developing a System of Indicators for Unmeasured Survey Quality Components. *Bulletin of the International Statistical Institute* 48: 469–486.

Mardia, K.V., J.T. Kent, and J.M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

McCulloch, C.E., and S.R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 141–142, http://dx.doi.org/10.1515/JOS-2015-0008

Book Review

Dean M. Resnick[1]

Anders Wallgren and Britt Wallgren. (Eds.) *Register-based Statistics*. 2014 New York: Wiley, ISBN 978-1-119-94213-9, 320 pp, $120.

"Register-based Statistics" by Anders and Britt Wallgren is a how-to cookbook for creating a national statistical register from scratch. The type of register envisioned is one along a Nordic model that continuously tracks a set of entities such as persons, households, or businesses by the compilation and updating of existing data from administrative sources. Created in this manner, this kind of register would allow the development of consistent, policy-relevant statistics on an ongoing basis or as new research questions arise without having to field a new survey, add new questions to an existing survey, or requiring the recompilation and reintegration of administrative record data from multiple sources. Based on the authors' experience of developing registers like this for Sweden, the authors provide tools, recommendations, caveats, and the rudiments of an administrative data system theory (which they correctly suggest is presently much less developed than sampling or survey theory).

For an American reviewer, this book presents something of a conundrum. This is because, at least in terms of person, family, or household-specific data, the development of a statistical register is not countenanced legally or socially, particularly under the coordination of a government entity. To some degree, this concern is obviated by the book's coverage of non-person-based registers (as of businesses), but more generally, the book takes on more relevance (for an American reader) if considered more as a guide to the use of administrative record data as combined from multiple sources, including survey data.

Here, this book provides a useful overview of the technical issues encountered in this type of processing (i.e., combining data from multiple sources). However, in this regard, this book should be considered more as an introductory presentation rather than a thorough explication of the more advanced data-management and statistical techniques needed for this. For example, the book discusses issues related to record linkage, imputation, entity duplication, and undercoverage, but in regard to these topics, a reasonably experienced data analyst or statistician would probably be seeking a much fuller treatment. Thus, it seems, this book is best suited for someone fairly new to the field, such as a manager or a policymaker. Here, the book lays out some very useful guiding principles, such as the need for subject-matter expertise, comprehensive metadata, and carefully thought-out data integration approaches.

[1] Health Policy Center, The Urban Institute. 2100 M Street NW Washington, DC 20037, U.S.A.
Email: DResnick@urban.org

Certainly, there are some areas that would be more valuable to a more experienced analyst. Particularly appreciated is the extensive treatment of multilevel variables. By this, the book means a categorical data item for which a given entity (e.g., a business, person or household) can be fairly considered as having more than one value, at least over the course of time. Here the authors rightly indicate the dangers associated with the selection and representation of only one of these values, such as the biasing of derived estimates, and provide thorough guidance on how multiple-level data can be retained and used for estimation. The recommended treatment of these data seems quite extendable to imputation results (although it is not clear this is intended by the authors).

In addition, this book provides a nice treatment of the integration of administrative and survey data, suggesting that some entities (i.e., businesses or households) may be represented on one of these sources and not another and therefore their concatenation allows a fuller picture of the represented situation than either alone. This would be advice well heeded for someone working to develop comprehensive statistical estimates from available data sources.

In terms of the treatment of error within administrative data, this book certainly provides good guidance on how to minimize these, but it is rather rudimentary in presenting a theoretical framework for quantifying them – suggesting the appropriate measurement techniques are not well developed. Here, the statistical comparison of data elements from different sources (i.e., comparing administrative data to survey data) seems a useful area for exploration and a natural extension of the treatment of data integration. Still, it is greatly appreciated that the authors stress that sampling error is only a small part of estimation errors (albeit readily treatable by known statistical techniques). If quality comparisons are made between survey and administrative data, the existence of nonsampling error in survey data should be recognized.

In terms of readers for whom this book would be most helpful, obviously, someone newly assigned to the task of creating a statistical register would be the greatest beneficiary. To some degree, persons with experience in this area would also benefit from the identification and systematization of methods relevant for this type of work. For those not involved in register-development *per se*, but seeking to develop competence in the integration and use of administrative record data, this book could be a useful introduction to and reference for applicable methods and their systematization and a source of best-practice principles.

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 143–145, http://dx.doi.org/10.1515/JOS-2015-0009

Book Review

Gina K. Walejko[1]

Frauke Kreuter. *Improving Surveys with Paradata: Analytic Uses of Process Information*. 2014 New York: Wiley, ISBN 978-0-470-90541-8, 416 pp, $74.95.

Since Mick Couper coined the term "paradata" in a presentation given at the 1998 Joint Statistical Meeting, the collection and use of paradata have expanded steadily. In this evolving environment, the edited book *Improving Surveys with Paradata: Analytic Uses of Process Information* insightfully contributes to the growing discussion on the advantages and challenges of using paradata.

Although the definition of paradata varies with each chapter's author, the book's editor, Frauke Kreuter, takes an inclusive view, defining paradata as "additional data that can be captured during the process of producing a survey statistic." (p. 3) Illustrating this broad definition, chapter authors discuss a range of paradata across multiple survey modes. For example, some investigate call-history data produced during computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI) contact attempts, which may include timestamps and attempt-level disposition codes. Others write about interviewer observations of housing units and sampled persons, for example, access impediments recorded during contact attempts and interviewer-documented household attributes related to key estimates such as the presence of a wheelchair ramp for health surveys. Others examine self-reported survey mode paradata including, but not limited to, questionnaire navigation data available from some web surveys that can reproduce a respondent's entire survey experience by recording mouse clicks and position, keystrokes, scrolling, page navigation, and timestamps. Such a comprehensive definition gives the fifteen-chapter book freedom to cover a variety of topics across the planning, data collection, and post-survey adjustment and analysis phases of the survey lifecycle.

Kreuter groups the book's chapters into three parts. Part One, Paradata in Survey Errors, applies the Total Survey Error framework as an organizing approach to discuss particular uses of paradata. Kreuter and Olson briefly examine the general concept of nonresponse bias and then explain how paradata have been used to identify nonresponse bias and perform nonresponse bias adjustments. The next two chapters similarly illustrate the use of paradata as they relate to measurement error summarizing the concept of measurement error in general. Olson and Parkhurst detail types of paradata produced across survey modes, while Yan and Olson briefly review studies that used paradata to investigate measurement error, giving four empirical examples. Eckman focuses on coverage error,

[1] U.S. Census Bureau, Center for Survey Measurement, 4600 Silver Hill Road, Washington, DC, 20233, U.S.A. Email: gina.k.walejko@census.gov

introducing readers to the concepts of undercoverage and overcoverage, and then explores how paradata can be used to uncover coverage bias across stages of frame construction.

Paradata in Survey Production, the second part of the book, not only is valuable in highlighting applications of paradata in surveys but also provides useful information on a variety of timely topics, for example, responsive design (Chapter 6), modeling best contact time (Chapter 7), within-survey requests such as consent for record linkage (Chapter 8), control charts and other quality control displays (Chapter 9), and representivity indicators (Chapter 10). Kirgis and Lepkowski introduce readers to the redesign of the 2006-2010 National Survey of Family Growth, focusing on five design changes that relied on paradata. Wagner illustrates the use of paradata-driven models to predict the best time of day to contact respondents in two surveys. Sakshaug outlines how paradata could increase response rates to four types of within-survey requests, including administrative record linkage, biomeasure collection, data-collection mode switching, and requesting sensitive information. Jans, Sirkis, and Morgan examine how survey managers can use paradata-based statistical quality control displays to manage survey performance. Schouten and Calinescu describe how paradata can be used to monitor contact, participation, and measurement "profiles" (i.e., classes of respondents that may be prone to measurement error), using the Dutch Labour Force Survey to show how administrative record data reveals measurement profiles associated with increased social desirability and satisficing behavior.

Part Three of the book, Special Challenges, includes five chapters dedicated to techniques for which the uses of paradata are not clear or may be challenging to utilize. Callegaro discusses device type, questionnaire navigation, and online panel web survey paradata, ending with the challenges of using such data, including privacy considerations and level of aggregation after collection. Durrant, D'Arrigo, and Müller give an overview of several multilevel modeling approaches that utilize call record data as model inputs, using two survey datasets to illustrate research questions these models could answer. Schafer describes how a Bayesian penalized-spline modeling approach can be used in statistical process modeling with paradata, thus allowing process means to vary over time. West and Sinibaldi perform a review of paradata quality, including an examination of mechanisms that may lead to errors in computer-generated and interviewer-observed paradata, and, finally, West presents the simulated results of weighting class adjustments when error levels of paradata vary.

The book's success can be attributed to the description of paradata and their uses in survey design, implementation, and analysis, and also to the care taken to clarify particular concepts. In addition to careful explanations of nonresponse (Chapter 2), measurement (Chapters 3 and 4), and coverage errors (Chapter 5), other chapters offer background information on survey and statistical concepts in the book. For example, Jans and colleagues discuss the history of control charts, the basic components of graphical displays, and rules for determining whether a subgroup mean is out of control (Chapter 9). Schafer devotes a large portion of his chapter to reviewing the uses of splines and showing how a penalized spline can be treated as a linear mixed model (Chapter 13). Although not related to paradata directly, the detailed overview of such concepts make chapters useful to both paradata newcomers and to experts looking to apply techniques explained in the book.

Although the book presents problems associated with the collection, analysis, and use of certain types of paradata, it offers a myriad of helpful suggestions for how to answer questions generated by these problems. Eckman encourages researchers to look at coverage bias in addition to coverage rates, stating: "Paradata can and should play an important role in this transition" (Chapter 5, p. 15). Wagner suggests several avenues for future investigations including optimal trip planning for face-to-face interviewers that incorporates clustered cases (Chapter 7), and West and Sinibaldi conclude that the entire chapter warrants additional evaluations of paradata quality (Chapter 14).

Improving Surveys with Paradata: Analytic Uses of Process Information adds to a list of excellent titles in the Wiley Series in Survey Methodology. The combination of teaching survey and statistical concepts with cutting-edge uses of paradata and challenges associated with such applications positions the book as a valuable resource for a broad audience, from students of survey methodology looking for a thesis project to seasoned survey practitioners solving a particular survey problem to veteran researchers analyzing paradata across multiple modes and studies. Although the applications of paradata will continue to evolve over time, the information presented in this book's chapters provides evidence of paradata's usefulness and persistence in the improvement of surveys.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 147–148, http://dx.doi.org/10.1515/JOS-2015-0010

Book Review

Gordon Willis[1]

Nick Emmel. *Sampling and Choosing Cases in Qualitative Research*. 2013 London: Sage Publications, ISBN 978-0-857025098, 192 pp, $125.

The selection of cases to study in qualitative research – that is, who to select, how to select them, how many to choose, and so on – may seem like an esoteric or niche area of research. These challenges, however, have become increasingly relevant in some areas of importance to survey researchers, such as my own discipline of the cognitive testing of survey questionnaires. Therefore, a practical book focusing on selection of participants for qualitative research could be extremely helpful, and Nick Emmel's recent contribution *Sampling and Choosing Cases for Qualitative Research* deserves consideration in this regard. The application to survey methodology is certainly not direct. To survey methodologists, Emmel would be considered an outsider; as a sociologist steeped in the traditions of qualitative research, he does not directly address the area of survey methods, or of selection of cases for qualitative endeavors within that science. However, as there is benefit in seeking input, perspective, and sources of new understanding from outside our usual sources, it is worth considering the lessons that might be gleaned from this work.

My overall conclusion is that the book will be of most use to survey researchers who are already well versed in the terminology, theoretical perspectives, and orientations represented by qualitative research traditions such as Grounded Theory, and who seek to expand the sophistication of their mastery with respect to sample selection for qualitative activities such as focus groups and cognitive interviews. The book is less appropriate for the survey researcher trained in cognitive psychology or statistics, as the author assumes considerable familiarity with the principles, terminology (jargon), and history of topics such as Grounded Theory, positivist versus constructivist philosophy, the Constant Comparison Method, and hermeneutic analysis. Similarly, the author assumes that readers already have knowledge of terms such as open, axial, and discriminate coding, and does not define these. Those of us not directly trained in the qualitative research tradition will therefore require an auxiliary glossary of terms to understand the arguments being expressed.

Even for readers who have already made an attempt to become educated in the qualitative research tradition, some of the material is very tough to negotiate. I get the impression that Dr. Emmel is an authority in the general discipline of qualitative research, whose ultimate desire in writing this volume was to break out of the chains imposed by the nominal topic of 'sampling and choosing cases' – and to tackle significant epistemological debates that have circulated throughout the qualitative research world. For instance, there

[1] National Cancer Institute, National Institutes of Health, Room 3E358, 9609 Medical Center Drive, Rm 3E358, MSC 9762, Bethesda, MD 20892-9762. Email: willisg@mail.nih.gov

is considerable discussion of the degree to which Grounded Theory approaches to theory discovery should be represented by the original, *tabula rasa* view, versus a later perspective that relies more on investigator contribution to the initial level of theorizing. Such debates are of course germane to sample selection, but address a much broader world, and would likely be most accessible to that subculture of theorists already engaged in these debates.

Still, there is considerable value for those who make use of qualitative methods in the survey field. Most significantly, Emmel's discussion makes clear that extensive debate exists within the qualitative research field concerning key approaches to the analysis of qualitative data. This insight serves us well as a protective barrier to the erroneous notion that there are ready answers to the challenges we face, if only we lose our disciplinary blinders and accept the truths embedded within a related, mature field. In fact, questions that bedevil cognitive interviewers regarding case selection – how many interviews to conduct, how to choose who to interview, and how to use results to in turn select more cases – are certainly not settled science within the more general qualitative literature. At the least, it is reassuring to discover that there is no convenient solution that we have been ignoring all along.

Although Emmel's approach is highly theoretical rather than practical, and in no sense provides a recipe book for the selection of cases in qualitative research, he does emphasize what he labels the "Realist approach" which takes into account resource constraints and the need to conduct work that is convincing as well as theoretically supported. To this end, he presents ideas that are useful to survey researchers, the most intriguing of which may be his analysis of saturation as a means for establishing overall study sample size. Although it is sometimes suggested that an obvious practice is to determine sample size by stopping when we have achieved saturation, he makes a good case that this is a somewhat nebulous objective. Although the criterion of 'testing until no new categories or findings are discovered' sounds clear enough, in practice the determination of exactly when and how such a state is achieved may vary widely, and is dependent on factors such as the level of effort put into maximizing variation in the sample and the extensiveness of coding or preliminary analysis. In application to the conduct of cognitive and other survey pretesting, an implication of Emmel's message appears to be that statements such as 'testing was done until saturation was achieved' are difficult to evaluate, and might even reflect an element of gaming the system (akin to questionable practices well known to survey researchers, such as presenting a Response Rate that is more accurately described as a Cooperation Rate).

Apart from its direct relevance to survey research, the book does provide some very interesting, informative, and thought-provoking examples and illustrations, invoking themes that include Guy Fawkes, Russian matryoshka dolls, and John Snow's investigation of cholera in Soho, London. The Snow example is particularly salient, as Emmel identifies the usual view of this, as an application of geographical mapping of cases that ultimately led to the identification of the Broad Street pump as the disease source, to be something of a scientific urban legend. The ultimate lesson that Emmel conveys is that explanations stemming from qualitative research are complex, and that we need to be very careful in deciding who, and what, we make use of, along the winding road traveled by qualitative researchers.