JOS

Journal of Official Statistics

Statistiska centralbyrån Statistics Sweden

An International Quarterly Published by SCB, Statistics Sweden

www.jos.nu

September 2013

Journal of Official Statistics
vol. 32, i. 1 (2016)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 1–28, http://dx.doi.org/10.1515/JOS-2016-0001

Micro- and Macrodata: a Comparison of the Household Finance and Consumption Survey with Financial Accounts in Austria

Michael Andreasch[1] *and Peter Lindner*[2]

This article compares the results of Austria's Household Finance and Consumption Survey (HFCS) on savings deposits and estimates on total financial assets with administrative records from the national accounts for the household sector. The microdata that are newly generated through the HFCS and the detailed (internally available) breakdown of savings deposits in the existing macrodata (financial accounts) lend themselves to a more in-depth analysis of the similarities and differences in these two sources. Comparing the data shows that the HFCS-based aggregate estimates are lower than the financial accounts data, which is in line with evidence from the literature. The article also shows, however, that the survey adequately captures the underlying patterns at the microlevel in terms of the overall financial portfolio allocation and the distribution of savings deposits over detailed breakdowns. Moreover, a simulation based on the HFCS data demonstrates the effect that the inclusion of savings deposits in the most affluent tail of the distribution has on common statistics. Undercoverage above all of the upper deposit ranges suggests an underestimation or bias in the statistics. This underestimation, however, can be shown to be relatively minor, particularly in the case of robust statistical measures, such as the median or percentile ratios.

Key words: Wealth distribution; survey; national accounts.

1. Introduction

In recent years, survey data have become an important tool in the research on assets and debt. The data often constitute the only pool of data on household assets that is collected systematically at the microlevel. Yet the tradition of surveys on household assets is shorter than that of income surveys. For this reason, survey data on incomes have been compared with income data from other sources more frequently and in greater detail in the literature. The innovation of the Household Finance and Consumption Survey (HFCS), which covers the entire eurozone, is that it provides a harmonised framework for collecting information on eurozone household (financial and nonfinancial) assets and liabilities, which represents a basis for eurozone-wide analyses.

[1] Oesterreichische Nationalbank, External Statistics, Financial Accounts and Monetary, and Financial Statistics Division. Otto-Wagner-Platz 3, 1090 Vienna, Austria. Email: michael.andreasch@oenb.at
[2] Oesterreichische Nationalbank, Economic Analysis Division, Otto-Wagner-Platz 3, 1090 Vienna, Austria. Email: peter.lindner@oenb.at

Acknowledgments: We would like to thank Nicolás Albacete, Pirmin Fessler, Martin Schürz and four anonymous referees for very helpful discussions and comments. In addition to the usual disclaimer, the opinions expressed in this study solely represent those of the authors and do not necessarily reflect the official viewpoint of the Oesterreichische Nationalbank or of the Eurosystem.

Although all forms of data compilation come with their own specific problems, some difficulties attached to surveys attract special criticism, such as nonparticipation or nonresponse. A key criticism is that households often decline to participate in voluntary surveys or that, if they do agree to participate, they provide incorrect information or refuse to respond to specific questions. In addition, the survey methods may influence results from survey data, for example, the interview mode (see Fessler et al. 2012). Hence, to identify the strengths and possible weaknesses of the HFCS data, it is useful to compare them thoroughly with other national statistics. In doing so, we also need to bear in mind that the macrodata exhibit certain weaknesses. The most obvious one is that data from financial accounts are (publicly) only available at the aggregate level and thus it is not possible to carry out a distributional analysis. Additionally, there are also issues concerning classification of the data (households vs. self-employed businesses/ other institutions) and estimations (e.g., cash holdings). Thus it is far from clear that one or the other source of data present a better choice for all investigations, and so comparing the results of the HFCS survey with other national statistics will contribute to a better understanding of the economy, as different data sources tend to generate complementary findings.

Furthermore, in the light of the "Report by the Commission on the Measurement of Economic Performance and Social Progress" by Stiglitz et al. (2009), which recommended to "[g]ive more prominence to the distribution of income, consumption and wealth" (Recommendation 4 on page 13), our understanding of the integration of micro- and macrodata must be analysed and enhanced. This analysis also contributes to the effort of international institutions such as the ECB to integrate information from the macro- and the microlevel to a greater extent. Furthermore, in light of the "Beyond GDP" initiative of the European Commission, the analysis at hand can be viewed as a first step towards an approach integrating micro- and macrostatistics. Before a clear view of the overall picture can be gained, we need to understand the similarities and differences between the existing information in detail.

One of the general results documented here is evidence that the HFCS in Austria underrepresents households' financial assets: total financial assets as identified by the HFCS come to roughly 40% of total financial assets as shown by the financial accounts (Section 4.1). Essentially, this finding corresponds to similar comparisons of survey data and administrative records described in the literature (Section 2). Owing to the internal availability of administrative records on financial wealth, the article contributes to the existing literature in the following ways. First, we compare the allocation of savings over different deposit ranges and different sectors of the Austrian banking system, as these are recorded by both the HFCS and existing national statistics (hence the article goes beyond a comparison of the aggregate statistics). We find that the deposit patterns are similar in both the survey data and the banks' reports. Furthermore, a microsimulation of the upper deposit amounts, which are underrepresented in the HFCS, shows that the ensuing (negative) bias is relatively low for statistical robust estimates in particular. Thus, depending on the issue under research, both the aggregated data of the national accounts and the HFCS data represent a valid basis for empirical evaluations. The results presented in this analysis should provide a good understanding of the relationship of the micro- and macrodata of other eurozone countries due to the harmonised manner of data collection and the similar relative importance of the major components.

This article is structured as follows. In Section 2, we establish a link between the article and the existing literature. Section 3 provides an explanation of the data used. The results of the comparison are presented in Section 4. In addition to the evaluation of aggregate results, we provide a comparison of the HFCS data with the banking statistics in a detailed breakdown of deposits on savings accounts. The simulation of the upper savings deposit ranges along with the evaluation of the impact of undercoverage on the main estimators is set out in Section 5. The analysis concludes with final remarks and suggestions for further research.

2. Background

Comparisons of survey data with data derived from administrative sources are common in the scientific literature. As data on flows of the household balance sheet, in particular from administrative sources, are more readily available than data on household stocks, most studies limit themselves to evaluating information on incomes. The literature comparing income in survey and administrative data is able to provide a broader picture of relevant ideas for investigations concerning the stocks of the household balance sheet, such as the present article.

In summary, income data from both survey and administrative sources are subject to errors, the resulting bias of the estimators is expected to be low, and, in most studies, the differences between the data result from specification differences (definitions of the unit of collection, of types of income, etc.). As a case in point, Törmälehto (2011) compares the data collected by the Luxembourg Income Study Group (LIS) with income aggregates in the national accounts. He observes that surveys capture over 90% of income in most countries, admittedly with a lower degree of coverage in some income subcategories. For the United States, Davies and Fisher (2009) find some differences between individual income sources using data from the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP) matched with administrative data from the social security administration. Using the same datasets, Roemer (2002) shows that the surveys accurately capture the underlying patterns of income distribution. Roemer also points out the problems underlying income distributions based on administrative data (e.g., because illegal work and related income are not captured in the administrative data). Kavonius and Törmälehto (2003) compare income aggregates of various sources from survey data (e.g. Income Distribution Survey) with national accounts data for Finland. While wages and salaries are nearly identical in both data sources (survey coverage is about 99%), the data for property income and self-employment income differ substantially (unadjusted coverage is 210% and 52%, respectively). Bricker and Engelhardt (2008) report on measurement error in earnings data for men and for women in the United States, comparing administrative records of the Social Security Administration (SSA) and of the Internal Revenue Service (IRS) with the survey data in the Health and Retirement Study (HRS). As the data can be precisely matched, the authors are able to identify a measurement error of about six percent in men's incomes and of approximately seven percent in women's incomes. Finally, Kapteyn and Ypma

(2007) research measurement error on the basis of data from the Swedish Longitudinal Individual Data Base (LINDA) compared with information from the Survey of Health, Ageing, and Retirement in Europe (SHARE). The authors show that erroneous observations lead to biased estimators in a variance analysis. Errors are found not just in survey data, but also in the administrative data.

The literature has not produced as many findings on stocks of the household balance sheet. Avery et al. (1988) were the first to compare aggregate estimates based on survey data with national accounts data (i.e., flow-of-funds statistics). The authors show that aggregate savings deposits as documented by the Survey of Consumer Finance (SCF) amounted to less than 50% of aggregate savings deposits as captured by the flow-of-funds statistics. However, the discrepancy between the two data sources with regard to the household wealth held in the households' main residence offsets this difference. Thus the estimate of households' gross assets is quite similar in both data sources. Similarly, Antoniewicz et al. (2005) examined the coverage of financial assets and liabilities of the household sector in three surveys performed for Italy, the United States, and to some extent for Canada. With regard to Canada, where data were available for 1999, the microdata on deposits and total liabilities were around 30% lower than the macrodata. This result is echoed by the microdata for Italy, which are based on the Survey on Household Income and Wealth (SHIW): the estimate for total financial assets in the SHIW came to 31% of the corresponding macrodata. However, an adjustment for underestimation and nonresponse produces a significant improvement of underreporting. In the United States, the survey data (SCF) are closer to the flow-of-funds data. In a more recent paper based on the same data, Henriques and Hsu (2014) show additionally that the changes in the aggregate values over time are broadly synchronized. Sierminska et al. (2006) compare the data of the Luxembourg Wealth Study (LWS) for several countries with national statistics. The authors show that the varied sources on which the LWS database is based capture between 13% and 117% of per capita household wealth. The administrative data are subject to some problems, so that an estimate of per capita household wealth in the LWS database equalling 117% of the estimate based on national statistics is not necessarily a sign of a lack of quality of the surveys used. With a ratio of the LWS database to the national balance sheet of between 65% and 117%, the match between the micro- and macrodata of nonfinancial assets is closer than that of financial assets (with an LWS to NBS ratio of between 13% and 52%). Finally, Johansson and Klevmarken (2007) used information from the administrative LINDA database and from two surveys conducted in Sweden (both refer to residents aged 50 and over) to identify measurement error, its correlation with the volume of assets, and the effects on regression analyses. The authors concluded that measurement error correlated with the volume of assets occurs above all at the tails of the distribution. In an independent effort at approximately the same time as this article was written, Kavonius and Honkkila (2013) looked at the comparison of the HFCS with National Accounts for Finland, Italy and the Netherlands. However, Kavonius and Honkkila (2013) only look at a comparison of aggregated values. The analysis below extends the literature by looking at detailed categories in terms of asset ranges and banking sectors on the one hand, and by simulating the potential impact of the highest saving levels on commonly used statistics on the other hand.

3. Data and Definitions

This analysis is based on two different datasets from Austria, data derived from the HFCS and administrative banking statistics used to compile the financial accounts. Both types of data are compiled and managed by the Oesterreichische Nationalbank (OeNB). The breakdown in both the microdata and macrodata permits a granular analysis of the interlinkages. Appendix A provides the details of the breakdown by banking sector and assets ranges.

3.1. The HFCS in Austria

The first wave of the HFCS is the most comprehensive survey on household assets and debt to be conducted in Austria. Of a stratified cluster random sample of 4,436 households, 2,380 households agreed to participate in the voluntary survey and were interviewed in person (CAPI - Computer-assisted personal interviewing) about the different components of household assets and liabilities among other things. The field phase was conducted from the third quarter of 2010 to the second quarter of 2011. The reference period for stock information is the time of the interview. Most of the missing information (i.e. information not provided by respondents) was imputed using a Bayesian-based multiple-imputation procedure (this is explained in more detail below). On the basis of sample design weights and after nonresponse adjustment, the final household weights used in the evaluations in this analysis were poststratified both by regional distribution of the households and by distribution of household size (see Albacete et al. (2012) and Fessler et al. (2012)). In particular, this means that the weights were not adjusted to meet the aggregates or the structure of wealth and debt positions of an administrative data source. Hence, differences between the two separate data sources are to be expected; they have not been reduced or ruled out ex ante in the production process.

3.2. The Financial Accounts in Austria

The financial accounts are an integral part of the national accounts and as such are compiled in accordance with the rules of the European System of National and Regional Accounts 2010 (ESA 2010) based on data derived from a variety of administrative sources. In particular, the following components are used for the compilation of the data on deposits:

- The OeNB’s financial statements,
- MFI (monetary financial institution) balance sheet statistics,
- supervisory statistics of banks resident in Austria,
- quarterly/annual balance of payments and international investment position data.

We used the financial accounts data for the reporting date 31 December 2010 (i.e., in the middle of the field phase of the HFCS) for comparison with the HFCS results. The focus of our analysis is not just on establishing the discrepancies between the aggregate values – as documented in the international survey literature – but above all on assessing the allocation of deposits to small ranges of volume and to the different sectors of the Austrian banking system. These data from the banking statistics are an important component of the financial accounts. This approach allows for the documentation of new and more detailed findings on the similarities and differences between macro- and microdata.

3.3. *Definition of the Unit of Collection*

The household represents the unit of collection in the HFCS. All households in Austria (except institutionalized households living, for example, in a home for the elderly, a monastery, military compound, or prison) are part of the target population, irrespective of their nationality, and thus have a positive probability of being selected for the HFCS sample.

By contrast, the banking statistics in the financial accounts capture the information on (euro-denominated) savings accounts, not by households but by accounts. These accounts can be allocated to the sector of (domestic) households and self-employed persons. The reports cover the accounts of all Austrian residents (persons or institutional units). The household sector includes consumer households, self-employed persons and sole proprietorships. Financial assets and liabilities for the self-employed businesses are shown on a gross basis in the financial accounts. In the HFCS, wealth of self-employed persons and sole proprietorships is classified as net wealth in self-employment business, that is, total assets (real and financial) minus liabilities, and is not recorded as part of the financial wealth but rather as real assets.

Household level in the survey and deposit account in the banking statistics are obviously two different units of observation. Despite the fact that it is the only possible way to compare savings from the two sources in detail as is done in this analysis, there are further reasons why this distinction does not render the analysis meaningless. As will be shown below, households have more than one account, but most households only use one bank, so the categorisation into banking sector is not affected to a large extent by the unit of observation. Furthermore, although shifts in asset ranges to higher ones might be expected in the survey due to aggregation of accounts, we would argue that the comparison of the detailed ranges is still valuable, since a lot of findings such as missing information in some ranges in the survey still provide important information independent of the discrepancy of the unit of observation. One can estimate how much is missing solely because of ranges with no observations in the survey, for example. Furthermore, bearing the unit of observation in mind allows us to see whether the aggregation at the household level yields the expected results, such as higher average values.

4. Results of the Comparison of HFCS and Financial Accounts Data

4.1. *Aggregates*

Major aggregate components of financial assets classified in the financial accounts can be estimated from the HFCS as well. The definitions of the information collected in the HFCS and reflected in the macrostatistics of the financial accounts are broadly comparable. Kavonius and Törmälehto (2010) have documented the link between the HFCS variables and the ESA definitions in detail, and so the links are not explained again. The following picture emerges for Austria (Table 1), with the top part of the table showing

Table 1. Comparison of HFCS and financial accounts aggregates

HFCS			Financial accounts				
	Total, € million	Share		Total, € million	Share	ESA	HFCS/financial accounts ratio
Sight accounts	11,847	7%	Sight accounts	47,878	11%	A F. 22	25%
Savings accounts (excluding life insurance)	60,287	35%	Savings accounts[1]	149,477	36%	AF22/AF.29	40%
			Other accounts[2]	7,161	2%		n.a.[2]
Bonds and other debt securities	13,635	7%	Money market paper	1,041	0%	A F. 31	32%
			Long-term debt securities	41,484	10%	A F. 32	
Shares, publicly traded	5,384	3%	Quoted shares	18,452	4%	AF. 511	29%
Net wealth in business, non self-employment and not publicly traded (HD1010)	2,249	1%	Unquoted equity	3,052	1%	AF .512	74%
Funds	20,852	12%	Mutual fund shares	41,509	10%	A F. 52	50%
Life insurance policies	38,571	22%	Life insurance policies	67,825	16%	AF. 62	57%
Pension wealth	20,531	12%	Pension fund reserves	31,515	8%	A F. 63	65%
Value of any other financial asset (HD1920)	1,650	1%	Other accounts including financial derivatives	9,813	2%	A F.7/A F.8	17%
Comparable financial assets	175,005	100%	Comparable financial assets	419,208	100%		42%
Components not contained in one of the data sources							
			Cash	16,853	3%	A F. 21	
			Loans	0	0%	AF.4	
			Investments in other equity[4]	78,849	15%	AF. 519	
			Nonlife insurance claims	9,612	2%	AF.61	
Debt to households	6,151	3%					
Managed accounts	5	0%					
Total financial assets[3]	181,161	100%	Total financial assets	524,522	100%		35%

Source: HFCS Austria 2010, OeNB; OeNB financial accounts (ESA 2010). For more details see: http://www.oenb.at/dms/oenbEN/Statistics/Download/Standardized-Tables/Financial-Accounts:

[1] In the financial acccounts, savings accounts also include non-euro savings accounts, which accounts for a small discrepancy between this aggregate and the comparable value in Subsection 4.2.

[2] The special item "other accounts" of total accounts in the financial accounts includes all time deposits that cannot be assigned to sight or savings accounts as well as savings abroad. To improve comparability, they were stated separately as part of savings deposits.

[3] This definition of financial assets includes household assets owned under occupational and private pension schemes; therefore, it is not in line with the definition the ECB uses in the HFCS.

[4] Other equity includes the equity in limited liability companies and imputed equity holdings in private foundations.

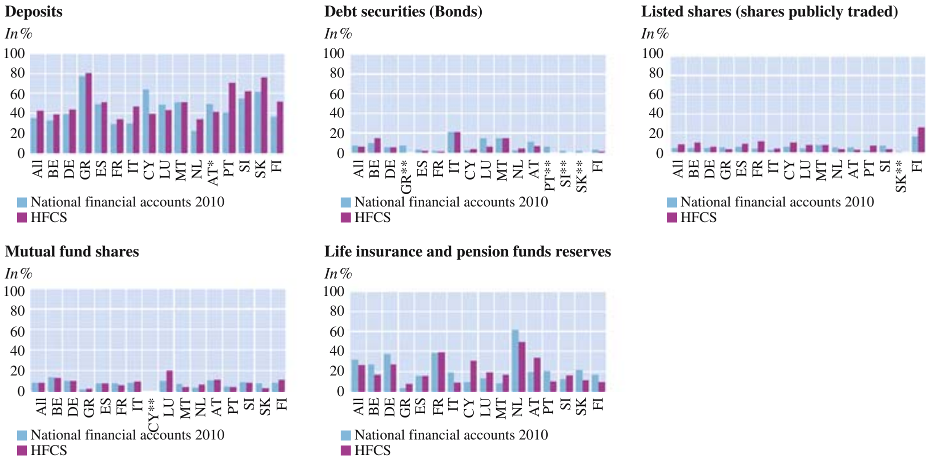
the comparable components, including the share of each component in terms of total comparable financial wealth and the bottom part showing the components that are not covered by one of the two data sources.

As common in the literature, the comparison of survey data (HFCS) and financial account aggregates indicates underreporting of household financial wealth in the HFCS in Austria. Table 1 indicates that the HFCS aggregate for comparable household financial wealth in Austria runs to about 42% of the financial accounts aggregate. This value may be considered fairly high in an international comparison with other surveys (see also Section 2). Sierminska et al. (2006) for example shows ratios ranging from 13% (United Kingdom, BHPS 2000) to 52% (Norway, IDS 2002) and Mathä et al. (2012) indicate a ratio of 35% for the HFCS in Luxembourg. The possible origins of this difference are manifold; on the one hand the survey estimates might not cover the totality of the financial assets, but on the other hand the financial accounts data do not reflect solely the financial wealth of households as they include self-employed business assets and single-person companies and thus overestimate the households' financial wealth. However, the table indicates that (i) the distribution of the individual components of comparable financial assets in the HFCS data broadly mirrors the financial accounts patterns (see columns headed "share") and that (ii) the coverage ratio of the HFCS compared with the financial accounts varies considerably for individual financial instruments and components (see column headed "HFCS/financial accounts ratio").

The HFCS/financial accounts coverage ratio for savings deposits runs to 40%. It must be noted, though, that the administrative records on total deposits also include the deposits of self-employed persons and sole proprietorships (accounting for €13 billion at the end of 2013), which the HFCS classifies as net investment in self-employment business, that is, as real assets. In the HFCS, the volume of life insurance holdings (representing the second-highest shares in both data sources) is calculated as the accumulated premia over the time span of the contract up to the time of the interview. The financial accounts data are based on insurance technical reserves comprising provisions for prepayments of premia (the difference between premia recognised and premia earned) and actuarial reserves (current value of expected future benefits); they may also include life insurance provisions if policyholders bear the investment risk. The HFCS captures premia, but no profit participation or service charges of the insurance providers. In addition, the value of life insurance holdings can fluctuate in the case of unit- and index-linked life insurance contracts.

Certain subcomponents are not covered by either of the two data sources. For instance, in the financial accounts, financial wealth resulting from the debt of a household to the respondent's household is not covered, as relevant data are not available. However, the HFCS shows that this component has a non-negligible volume. The HFCS did not include a question on cash holdings, as this question was considered to be too sensitive. In the financial accounts, the category "cash holdings" is calculated based on the estimated proportion of total financial assets adjusted by the change in cash requirements for consumption.

Figure 1 shows the relative importance of the major components of financial wealth for all countries covered in the HFCS. The similarity of the overall distribution of components of comparable financial wealth holds not only for Austria, but for all countries



Source: Eurostat, Eurosystem HFCS (ECB (2013), table 2.6, page 47).
Note: * In Austria wealth held in life insurance is excluded from deposits; In countries marked with ** the estimate for the portfolio item is suppressed due to less than 25 observations (see ECB (2013)).
Country abbreviations are as follows: Belgium (BE), Germany (DE), Greece (GR), Spain (ES), France (FR), Italy (IT), Cyprus (CY), Luxembourg (LU), Malta (MT), the Netherlands (NL), Austria (AT), Portugal (PT), Slovenia (SI), Slovakia (SK), and Finland (F I).

Fig. 1. Proportion of financial asset categories as share of total financial assets

participating in the HFCS across the board. Deposits (sight and savings accounts together) account for some 42% of financial wealth in the HFCS and some 47% in the financial accounts in Austria. Thus, these holdings make up the largest share of financial assets. Consequently, the analysis of this component of financial wealth has a greater explanatory weight (see Andreasch et al. (2009) for a comparison of survey data and administrative data on investments in selfemployment businesses). Given the fact that a breakdown of deposits compiled in macrodata by individual households is not possible, the attempt was made to find reasons for the discrepancies in the total volume by the analysis of data by individual banking sectors and asset ranges. This breakdown is available in both sets of data sources. In addition, the macrodata broken down by banking sectors are further disaggregated in different ranges of level of deposits, including the number of accounts allowing the estimation of the average amount for each range of deposit.

With some exceptions, the structural pattern in other countries seems to broadly reflect what is found in Austria. Hence – together with the ex-ante harmonisation of the HFCS – we are convinced that the remaining results in this study are a reasonable indication for other countries as well. In the following, we are able to extend the literature by making use of the detailed administrative records with respect to savings accounts.

4.2. Comparison of Savings Deposits

4.2.1. Historical Background and Imputations

In the Austrian financial landscape, savings accounts for a very long time enjoyed a special position, as depositors were able to hold numbered accounts and thus remain anonymous. Opening anonymous accounts has been prohibited by law since 2000; and since then customers have been required to provide identification when opening an account. In theory, it is still possible to hold anonymous accounts even today, as the requirement imposed on banks is to identify accounts only if there are withdrawals or payments into the account. Additionally, the identification of existing savings accounts is reported to the Austrian Federal Ministry of the Interior only for withdrawals from deposit accounts with an amount of above €15,000. The historical development of identification requirements for savings accounts and the tradition of keeping information about household wealth, especially savings, confidential – households consider this information personal and sensitive – explains households' reluctance to provide information on the volume of holdings in savings accounts in the survey.

Based on the flags which describe the origin of every observation and used for the variable for deposits on savings accounts (HD1210), Table 2 shows that (only) about 56% of respondent households provided the exact amount of holdings in savings accounts. Approximately four percent of households could not ("don't know") and about ten percent did not want to ("no answer") provide data. An additional 16% of households provided range estimates, as they were unable to indicate specific amounts. This shows that in a voluntary survey like the HFCS, not only unit nonresponse (refusal to participate) but also item nonresponse (refusal to answer particular questions) represents a difficulty, especially when questions cover such sensitive issues. As the (partial) lack of answers cannot be considered purely random, the exclusion of these households (commonly referred to as "listwise deletion" or "complete case analysis" in the literature) results in a distortion of the estimators. Thus, in line with the procedures applied in the recent literature, the missing information in the HFCS was imputed using Bayesian-based multiple imputation (see Albacete et al. 2012 for an in-depth explanation of the imputation procedure applied). The estimations in this study take the multiple-imputation structure and survey design into account.

4.2.2. Comparison of Number of Accounts

The banking statistics documented roughly 23.5 million savings accounts as of the end of 2010, and according to information provided by Statistics Austria, some 8.4 million persons (3.7 million households according to the HFCS estimate) live in Austria. Hence, many persons have several savings accounts, but the amounts held in these accounts are

Table 2. Share of imputed observations

	Number	Share
Not applicable (no value due to use of filter)	295	12.4%
Value collected, complete observation	1,321	55.5%
Edited, value collected was incorrect	2	0.1%
Imputed, originally – Don't know	83	3.5%
Imputed, originally – No answer	244	10.3%
Imputed, originally not collected due to higher order missing	38	1.6%
Imputed, originally collected from a range or from brackets	381	16.0%
Imputed, collected value deleted or value not collected due to CAPI error	16	0.7%
Total	2,380	100%

Source: HFCS Austria 2010, OeNB.

fairly small (see Table 3: roughly 81% of accounts contain deposits of less than €10,000). The reasons for having more than one savings account can be summed up as follows:

- Savings plans with building societies are separate savings accounts subject to special tax treatment. Therefore, many persons (Austrian citizens) have at least two savings accounts, one being a savings plan with a building society and the other a standard savings account. Customers typically attribute their building society savings plan to their house bank even though legally speaking, the deposits are held with another bank (a building society).
- Furthermore, security deposits for rental apartments are frequently kept on a separate savings account.
- As account maintenance charges are low (some Austrian banks do not charge any maintenance fees for accounts), people often have several savings accounts so that they can react quickly to interest rate differentials.
- Separate savings accounts (and partly also savings plans with building societies) are also kept for children.
- In addition, some account holders may have in fact forgotten they have accounts with very small holdings, so that the banking statistics may overrepresent actively held savings accounts. These forgotten accounts are by law kept alive for 30 years upon which they expire if no bank transfer (apart from interest payment) occurs in this period. Especially in the lowest deposit categories, the number of accounts may be distorted upward in the banking statistics in terms of active accounts.

Table 3 shows the distribution of the number of savings accounts by deposit holdings. The number of savings accounts is not explicitly asked for in the HFCS. However, the number of customer relationships households in Austria have with different banks can be estimated. The result of this calculation on the basis of HFCS data is displayed in the first column of Table 3, which indicates the number of customer relationships broken down by deposit ranges and the sum total of about 4.2 million of these relationships, which compares with about 23.5 million accounts in the financial accounts. Moreover, the table shows that the aggregation of potentially many accounts results in a higher percentage of

Table 3. Number of customer relationships with a bank/savings accounts

	HFCS		Banking statistics	
	Total	Share	Total	Share
All accounts	4,205,802	100.0%	23,463,618	100.0%
Up to €10,000	2,653,396	63.1%	19,058,885	81.2%
€10,000 to €20,000	637,071	15.1%	3,207,943	13.7%
€20,001 to €50,000	533,765	12.7%	798,045	3.4%
€50,001 to €100,000	212,675	5.1%	271,481	1.2%
€100,001 to €500,000	166,324	4.0%	119,911	0.5%
€500,001 to €1,000,000	2,570	0.1%	5,019	0.0%
€1,000,001 to €3,000,000	.[1]	.	1,963	0.0%
Over €3,000,000	.	.	371	0.0%

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

[1] All cells marked with "." have no observation.

customer relationships with higher deposits in the HFCS than in the banking statistics: some 81% of all accounts belong to the lowest category (holdings of up to €10,000) in the banking statistics, whereas only about 63% of the accounts captured by the HFCS have holdings in this range. This difference is then spread among the next highest categories. As the individual accounts in the banking statistics cannot be assigned to individual households, it cannot be determined whether the aggregation of accounts within a household explains the totality of the discrepancy.

The HFCS does not capture accounts with holdings above €1 million. Oversampling of wealthy households could improve the coverage of savings deposits in the HFCS. The probability of a household having savings deposits of over €1 million is highly unlikely, as only a total of about 0.03% of savings accounts are classified in the top three categories. Only about 0.0099% of savings accounts are classified in the top two categories in the banking statistics. Conversely, the HFCS covered a sufficient number of households with savings deposits of up to €500,000, and few households in the range in between.

4.2.3. Savings Deposits Aggregate

The total volume of savings deposits of domestic nonbanks in Austria is about €156 billion. The overwhelming majority (i.e., roughly €150 billion or 96%) of this total can be attributed to households in the financial accounts. The remaining part is classified as “others”. However, the total of the household sector as derived from the banking statistics cannot be broken down further into individual ranges and into banking sectors for the household sector. Therefore, the value of about €156 billion for total domestic nonbanks is used for the analysis, even though this leads to an overestimation on the side of the administrative data.

A detailed breakdown of the differences between HFCS and banking statistics data are shown in Table 4. In the first row, total savings deposits in all banking sectors are shown in the HFCS (panel 1) and in the banking statistics (panel 2). The third panel shows the HFCS to banking statistics ratio of each value. The HFCS results in the following tables are based on the information provided on savings deposits; this data is attributed to banking sectors on the basis of the bank at which a household holds the highest amount of deposits. The appendix contains equivalent tables based on national deposit variables.

The HFCS does not contain information about the two highest deposit categories. Consequently, assets in this part of the distribution are underestimated. The volume of savings deposits is also underestimated in the HFCS in the lower categories. For instance, in the savings deposit category €100,000 to €500,000, HFCS coverage comes to nearly 87% of the total aggregate, but to only 19% of total of savings deposits up to €10,000. This underestimation is attributable above all to the aggregation of savings accounts at the household level in the HFCS rather than the account level (banking statistics). This pattern is similar across all banking sectors. The higher estimate for the aggregate value (HFCS) in the middle savings deposit categories in the joint stock banking sector is also a consequence of the difference between unit of collection at the household and at the account level. The banking statistics data show a relatively larger number of deposit accounts among the lower deposit categories. These banking statistics data are not suited to showing the distribution of savings by households in Austria, only by accounts. In addition to what is already documented in the literature, we see in particular coverage rates in the different deposit categories and in the different banking sectors.

Table 4. Aggregate savings deposits in individual banking sectors and deposit range

HFCS – total deposits, € million

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	60,287	7,766	8,378	14,214	11,025	17,545	1,359	.[1]	.
Joint stock banks	18,135	2,048	2,173	4,067	2,898	6,947	.	.	.
Savings banks	14,360	1,744	2,389	4,137	2,897	2,564	628	.	.
Raiffeisen credit cooperatives	17,187	2,589	2,577	3,595	2,944	5,144	338	.	.
Volksbank credit cooperatives	4,214	605	529	957	978	1,144	.	.	.
State mortgage banks	1,587	185	155	127	198	700	222	.	.
Other	4,805	594	555	1,330	1,111	1,045	171	.	.

Banking statistics – total deposits, € million

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	156,217	40,859	43,431	24,667	18,425	20,180	3,308	3,004	2,345
Joint stock banks	39,032	11,168	15,100	3,931	3,051	3,504	651	740	887
Savings banks	41,490	10,935	10,709	6,308	4,973	5,536	1,058	1,086	885
Raiffeisen credit cooperatives	56,118	13,744	12,379	11,682	7,997	8,194	1,057	745	319
Volksbank credit cooperatives	13,724	4,137	3,747	1,783	1,561	1,817	318	227	134
State mortgage banks	5,765	861	1,476	944	827	1,115	219	204	120
Other	87	13	20	19	16	13	4	2	0

HFCS – deposits, share in % of banking statistics figure

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	38.59%	19.01%	19.29%	57.62%	59.84%	86.94%	41.09%	.	.
Joint stock banks	46.46%	18.34%	14.39%	103.46%	94.99%	198.23%	.	.	.
Savings banks	34.61%	15.95%	22.31%	65.58%	58.26%	46.31%	59.36%	.	.
Raiffeisen credit cooperatives	30.63%	18.84%	20.82%	30.77%	36.81%	62.78%	31.96%	.	.
Volksbank credit cooperatives	30.70%	14.62%	14.12%	53.68%	62.64%	62.98%	.	.	.
State mortgage banks	27.53%	21.48%	10.50%	13.46%	23.95%	62.79%	101.51%	.	.

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "." have no observation.

Table 5 additionally provides an analysis of the shares of individual banking sectors (left part) and deposit categories (right part) in total savings deposits.

The allocation of deposit holdings to the individual banking sectors is broadly the same in the HFCS and in the financial accounts. For instance, the smaller banking sectors (the Volksbank credit cooperatives and the mortgage banks) account for deposit shares of 7% and 3% according to HFCS data. The comparable banking statistics values are 9% and 4%, respectively. Both data sources also show the three banking sectors holding the higher market shares of deposits. Only joint stock banks are shown to have a lower share and Raiffeisen banks a somewhat higher share in total deposits in the banking statistics.

According to the banking statistics more than two-thirds (roughly 70% in total) of all savings deposits are in savings accounts with holdings of less than €50,000 (see the right half of Table 5). The HFCS column features larger percentages of deposit holdings in higher categories due to the aggregation at the household level. Thus more than two-thirds of total savings deposits (71%) are held in the categories spanning the range from €20,001 to €500,000. This is yet another area in which the household-level data from the survey complement the banking statistics data, as the preferred unit of evaluation is usually the household, not the individual account. Although deposits in the range from €500,001 to €1.000,000 account for 2% of the total volume in both data sources, the two top categories (four percent of the total volume in the banking statistics) are not covered in the HFCS. This means in particular that nearly seven percent of the total undercoverage in the HFCS can be attributed to the top two categories.

4.2.4. Accounts with MFIs/Customer Relationships with Banks in the HFCS

In order to explore further similarities and differences between the two data sources beyond the aggregates and aggregate shares, we analyse the allocation of customer relationships with banks in the HFCS and of the numbers of accounts in the banking statistics (see Table 6). The first row in the HFCS panel ("total") differs marginally from the results in Table 3, as the percentages cover only the customer relationship with the bank with the highest deposit holdings.

The distribution of customer relationships (HFCS) in the individual cells is very similar to the distribution in banking statistics. For example, 32.9% of accounts are held in the joint stock banking sector according to banking statistics, and 28.9% of households have accounts in the joint stock banking sector according to HFCS data. The gap in the Raiffeisen credit cooperative sector is even smaller at 30.5% (banking statistics) versus 30.2% (HFCS). A broad view of all categories in the individual sectors reveals that the middle categories in all sectors are somewhat overestimated, whereas the categories at the upper and lower ends are underestimated in the survey. We should point out that less than 1% of accounts as shown by the banking statistics are in the category from €100,001 to €500,000 and that the HFCS estimates for this category are generally also of the same order (with the exception of the category joint stock banks). Hence, the HFCS appears to cover the customer relationship patterns quite well up to a level of about €500,000.

According to the banking statistics, all categories above €500,000 contain a maximum of 0.01% of accounts across all banking sectors. The HFCS contains nearly no observations above the level of €500,000. These figures once again show how unlikely it is that (enough) households with savings deposit holdings in excess of €500,000 will be

Table 5. Allocation of deposits to banking sectors and deposit range

	HFCS	Banking statistics		HFCS	Banking statistics
Banking sectors			*Deposit ranges*		
Joint stock banks	30%	25%	up to €10,000	13%	26%
Savings banks	24%	27%	€10,001 to €20,000	14%	28%
Raiffeisen credit cooperatives	29%	36%	€20,001 to €50,000	24%	16%
Volksbank credit cooperatives	7%	9%	€50,001 to €100,000	18%	12%
State mortgage banks	3%	4%	€100,001 to €500,000	29%	13%
Other	8%	0%	€500,001 to €1,000,000	2%	2%
			€1,000,001 to €3,000,000	.[1]	2%
			over €3,000,000	.	2%

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "." have no observation.

Table 6. Allocation of customer relationships/accounts to banking sectors and deposit range

HFCS – share of customer relationships

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	100.0%	61.0%	17.3%	13.5%	4.9%	3.2%	0.1%	.[1]	.
Joint stock banks	28.9%	17.9%	4.8%	3.8%	1.3%	1.3%	.	.	.
Savings banks	23.5%	13.2%	4.8%	3.7%	1.2%	0.4%	.	.	.
Raiffeisen credit cooperatives	30.2%	19.1%	5.3%	3.6%	1.4%	0.9%	.	.	.
Volksbank credit cooperatives	7.1%	4.5%	1.0%	0.8%	0.5%	0.3%	.	.	.
State mortgage banks	1.9%	1.1%	0.3%	0.2%	0.1%	0.1%	.	.	.
Other	8.4%	5.2%	1.2%	1.4%	0.5%	0.2%	.	.	.

Banking statistics – share of accounts

	Total	Up to €10,000	€10,000 to €20,000	€20.001 to €50.000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	100.0%	81.2%	13.7%	3.4%	1.2%	0.5%	0.0%	0.0%	0.0%
Joint stock banks	32.9%	27.2%	4.9%	0.6%	0.2%	0.1%	0.0%	0.0%	0.0%
Savings banks	26.1%	21.3%	3.4%	0.9%	0.3%	0.1%	0.0%	0.0%	0.0%
Raiffeisen credit cooperatives	30.5%	24.4%	3.8%	1.6%	0.5%	0.2%	0.0%	0.0%	0.0%
Volksbank credit cooperatives	8.3%	6.7%	1.1%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%
State mortgage banks	2.3%	1.7%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "." have no observation.

obtained. Appropriate oversampling of more affluent households in the survey might increase the chance of capturing the right tail of the distribution.

4.2.5. Average Deposit Holdings in Banking Statistics/in the HFCS

As the banking statistics data show both the volume of deposits and the number of accounts, the average holdings per accounts can be calculated. The arithmetic mean of deposits in households including the standard error of the estimator can also be estimated on the basis of the survey data. Table 7 shows the average deposit holdings broken down by deposit ranges and banking sectors for both data sources. For the HFCS data, the calculation of the standard error of the respective mean in a cell is based on 1,000 resampling weights contained in the HFCS data. A rescaled bootstrap procedure is the replication method used to construct the replicate weights. For details on the construction and use of these weights, see Albacete et al. (2012). Although it would be desirable to compare the whole distribution (or at least also the median), such a comparison cannot be made, as the banking statistics lack the relevant information.

Table 7 highlights two important aspects, namely (i) the total average of deposit holdings (Column 2) is higher according to the HFCS data than according to the banking statistics, and (ii) amounts above €500,000 are not covered, a confirmation of the known finding. The higher means are the result of the aggregation of individual accounts to household deposit holdings in the HFCS. The table shows clearly that the average amount of deposits in an account does not correspond to the average of Austrian households' savings deposit holdings, as households may have several accounts.

In the individual categories covered by the HFCS, the mean value of both data sources is similar. As a case in point, the average holdings of deposits in the range from €100,001 to €500,000 come to about €168,000 according to HFCS data (the standard error is roughly €19,000), thus matching the banking statistics average of about €168,000. Only in the first category – deposits up to €10,000 (and to a much lesser extent in the second category as well) – are the averages according to the banking statistics data far lower than the corresponding HFCS values. Savings accounts with very low deposits are responsible for this discrepancy. No large differences across banking sectors are observed, as the data from both sources confirm.

5. Simulation of the Impact on some Key Indicators in the HFCS

Finally, a look at the theoretical impact of coverage of the top deposit categories in the HFCS on commonly used statistics is able to provide some insights. The following simple simulation makes it possible to quantitatively assess how some indicators would change if the HFCS sample contained households with savings in the two top categories (savings of over €1 million). The HFCS already includes observations – albeit very few – in the category with savings of €500,001 to €1,000,000. The procedure simulates a few households with average holdings in the top two categories as available from the banking statistics. These households are assigned a weight, and the distributional indicators are then calculated with and without these households. The details of each step are laid out in the following paragraphs.

The household simulation is performed on the basis of the following assumptions: Two households with average holdings of €6,320,000 (average in the highest deposit range

Table 7. Average deposit holdings by banking sectors and deposit range

HFCS – average deposit holdings, €	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	18,333	3,869	14,737	31,943	68,297	167,958	679,387	.[1]	.
(Std Err)	*1,751*	*107*	*184*	*529*	*2,190*	*10,648*	x[2]	.	.
Joint stock banks	19,070	3,488	15,091	32,131	68,050	166,741	.	.	.
(Std Err)	*2,582*	*197*	*353*	*1,263*	*3,296*	*18,881*	.	.	.
Savings banks	18,610	4,012	15,091	33,647	72,815	176,697	.	.	.
(Std Err)	*2,801*	*220*	*346*	*1,092*	*5,336*	*23,938*	.	.	.
Raiffeisen credit cooperatives	17,280	4,122	14,852	30,554	64,917	174,409	.	.	.
(Std Err)	*2,506*	*178*	*397*	*1,409*	*3,691*	*24,678*	.	.	.
Volksbank credit cooperatives	18,025	4,047	15,641	35,216	65,442	136,913	.	.	.
(Std Err)	*2,905*	*439*	*1,069*	*3,111*	*6,181*	*36,368*	.	.	.
State mortgage banks	25,942	4,965	14,771	24,420	60,936	161,124	.	.	.
(Std Err)	*12,610*	*694*	*1,256*	*4,302*	x	*55,421*	.	.	.
Other	17,387	3,493	14,120	29,461	70,794	179,427	518,000	.	.
(Std Err)	*2,987*	*303*	*751*	*1,403*	*5,457*	*64,323*	x	.	.

Banking statistics – average deposit holdings, €	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Alle	6,658	2,144	13,539	30,909	67,867	168,288	659,002	1,530,099	6,319,995
Joint stock banks	5,057	1,753	13,126	30,305	68,117	166,951	663,942	1,564,715	7,709,296
Savings banks	6,787	2,186	13,411	31,058	66,302	172,241	664,159	1,565,499	6,319,543
Raiffeisen credit cooperatives	7,852	2,405	14,029	30,726	67,828	163,988	645,182	1,440,230	4,906,600
Volksbank credit cooperatives	7,078	2,616	13,901	32,504	71,089	172,677	664,868	1,445,637	5,601,417
State mortgage banks	10,644	2,219	13,977	31,921	71,279	179,325	677,077	1,684,248	4,446,296
Other	17,556	4,944	13,805	31,903	70,347	184,056	848,000	1,660,000	.

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "." have no observation.

[2] In all cells marked with "x" standard errors cannot be estimated on account of the small number of observations.

in the banking statistics) and four households with average savings deposits of €1,530,000 (average in the second-highest deposit range in the banking statistics) are imputed. While the assumption of the number of households is ad hoc, it is justified for two reasons: first, the assumption reflects the higher number of accounts in the second-highest deposit range in the banking statistics, and second, it allows for the assignment of different weights to the households.

Assuming that every household in Austria has the same number of savings accounts, there are roughly 330 households with accounts in the second-highest category and only about 60 households with accounts in the highest category. Hence, the nonresponse-adjusted weights are assumed to be very low[3]; that is, for the households in the top deposit range, the weight is 175, or approximately the smallest nonresponse-adjusted weight in the original sample. For two households in the second-highest deposit range the nonresponse-adjusted weight is set to 300, or roughly the smallest percentile of these weights in the original sample. For the remaining two households in the second-highest category, this weight is set to 750, or roughly the fifth percentile in the original sample. To influence the preparation of the survey as little as possible, the HFCS poststratification process in Austria was repeated with these newly simulated households. This last step in defining the final household weights is based on the nonresponse-adjusted weights as well as information on household size and the geographical distribution of households in Austria. For the simulated households, the information on household size and geographical location required for the poststratification process are randomized (uniform distribution). This means that the simulated households are assigned a random size of between one and six members (this corresponds to the minimum and maximum numbers of adult members in the households represented in the HFCS) and are assigned randomly to an Austrian province. In the poststratification procedure, the weights of the new total of 2,386 households are adapted in line with the distribution of household size and geographical location in Austria as taken from the Statistics Austria microcensus (see the HFCS documentation for Austria in Albacete et al. (2012)). After poststratification, the weights of the simulated households average 423 (408 prior to poststratification), whereas all other households have an average weight of around 1,600. The range of the weights changes from 150–750 to 159.6–721.3, that is, the range becomes smaller.

This simulation procedure reflects the relatively low number of accounts in the two top categories in the banking statistics. However, assuming an even distribution of the accounts, the six simulated households with an average weight of over 400 tend to overrepresent the roughly 400 households cited above. Thus it must be assumed that the simulation results represent the upper limit of the possible change.

Some of the most widely used indicators of the new sample can be compared with the estimators of the sample without the imputed households (original sample). The results are shown in Table 8.

Unsurprisingly, aggregate total savings deposits in Austria and average savings deposits are higher in the simulated sample. While the increase by 9% is economically significant, it cannot fully explain the entire underrepresentation (see Table 1 in Section 4). However,

[3] Increasing these weights does not necessarily exert a clearly defined effect on the estimators, as the nonresponse-adjusted weights are poststratified.

Table 8. Simulation results

	HFCS		
	Original sample	Simulated sample	Change from original sample (%)
Mean (€)	18,333	19,974	8.9%
Median (€)	6,985	6,994	0.1%
Gini	0.681	0.706	3.7%
P90/P10	64.68	64.57	−0.2%
P90/P50	6.23	6.23	−0.1%
P10/P50	0.10	0.10	0.0%
Aggregate (€ million)	60,287	65,731	9.0%

Source: HFCS Austria 2010, OeNB.

the quality of the simulation is also reflected by the absolute rise by some €5 billion, so that the aggregate in the top two categories of the banking statistics is fully covered. The impact on robust statistics such as the median or the percentile ratios is very small: the median of savings deposit amounts rises by just 0.1%, for example. The impact on the ratios of the percentiles is also negligible in all parts of the distribution. The minimal reduction of P90/P10 and P90/P50 can be explained by the fact that the 90th percentiles increase less than the 10th and 50th percentiles on account of the simulation. Conversely, nonrobust statistics such as the Gini coefficient or the arithmetic mean of savings deposits change more strongly. Factoring in the simulated households causes the Gini coefficient to go up by some 2.5 points (about four percent of the rise in inequality as measured by the Gini coefficient). The reason for this fairly strong effect is the widening of the wealth bandwidth in deposits. In the original calculation, the Gini coefficient is calculated for a bandwidth of €0 to less than €1 million. The inclusion of the simulated households with holdings over €6 million has an effect on the Gini coefficient, even though these households have a low weight.

Overall, the simulation exercise shows that the HFCS is very well suited to capturing most of the distribution (see percentiles) even without generating information on the upper ranges of savings deposits. With respect to the other indicators, oversampling of the wealthy households – and thus achievement of a higher probability of capturing very high savings deposits – would be desirable, but the current indicators still deliver the best estimators for these statistics. Capturing the households with the holdings in the highest savings deposit ranges would, if anything, increase (but not decrease) the estimators for the aggregate, for the arithmetic mean, and for the inequality of savings deposits as measured by the Gini coefficient.

6. Concluding Remarks

This article examines the similarities and differences between data derived from surveys and from administrative sources, focusing on savings deposits as the main category of households' financial wealth in Austria. To this end, we compare the aggregate values, in line with the approach commonly described in the literature, and additionally compare a detailed breakdown of deposits by banking sectors and by deposit ranges, which has not been documented in the literature so far. Given the ex-ante harmonisation of the HFCS and the relatively similar structure of the relative importance of the components of financial wealth (see Figure 1), results are expected to be similar in other eurozone countries.

The main results of this analysis and what we can learn from them may be summarised as follows: the HFCS is well suited to identifying the (basic) deposit patterns, but estimates of total wealth are distorted downward, as has already been previously shown in the literature (and is discussed in Section 2). The underrepresentation of deposits across all banking sectors and deposit ranges and the lack of information on the highest deposit ranges are the reasons for this underestimation. Oversampling in the HFCS may contribute to closing this information gap at the tail of the distribution in the future (although due to the extremely low number of accounts in the highest ranges it is by no means guaranteed). The aggregate measures derived from administrative sources should provide a reliable estimator.

In addition, we consider the effects of the different units of aggregating savings deposits in the banking statistics (accounts based) and in the HFCS (household based). The banking statistics do not allow individual accounts to be allocated to households. The aggregation of accounts to the level of households, which is done the HFCS, results in a shift across deposit ranges. This shift indicates that even the data reported by the banks in the banking statistic cannot be used to analyse individual households, so that the HFCS provides highly useful additional information to the aggregates. Furthermore, the distribution across banking sectors and asset ranges of deposits is relatively similar in both data sources. Consequently, the two data sources are not meant to replace each other; much rather, they serve as complementary sources for analysing households in an economy where reliable distributional estimates can be calculated from the HFCS and aggregate values from the financial accounts. A final simulation of the top savings ranges indicates that the estimators (such as the Gini coefficient or the arithmetic mean) from the HFCS represent at least a lower bound for the true parameters, and that some indicators, in particular robust statistics such as the median and percentiles, are affected to a fairly low extent. The survey data provide a wealth of information that complement the administrative data and that are needed in particular to analyse certain groups of the respective target population.

Many other areas of the household accounts were not examined in this study, which focuses on financial assets and in particular savings deposits. Future research could be devoted to other components of financial wealth, such as equity wealth, or the debt side of the household balance sheet. A more in-depth comparison of data on real assets would also be desirable. However, very little useful administrative data on real assets is available. Furthermore, the investigation of measurement error that could not be achieved with the administrative records at hand should yield interesting insights.

Appendix A: Explanatory Note on Data and Definition

The data available allowed for a comparison not only of the aggregate values, but also of transferable deposits (F.22) and savings deposits (as a subcomponent of other deposits, F.29) in a particularly detailed way. Exploiting this detailed information from administrative sources provides the opportunity to extend the results in the literature, investigating financial assets, not only total values but also the distribution over asset ranges and banking sectors.

The HFCS in Austria includes one question on sight accounts and two sets of questions on savings accounts. First, households are asked to specify the total amount of their savings deposits, broken down by (i) savings other than savings with building societies and (ii) savings with building societies (Note that life insurance funds must be subtracted from variable HD1210 of the version of the HFCS in Austria published by the ECB (this variable covers savings accounts) to ensure comparability with the values in the financial accounts). Building societies are banking entities that collect savings, usually from individuals, and grant preferential mortgage loans. Second, households are asked to indicate which banks they use based on a predefined list of the largest 21 banks and an additional verbatim recording for other institutions (up to five banks could be reported) and to specify how much money they hold in savings accounts and custody accounts at these banks, starting with the bank at which they hold the highest amount. The data from the first survey method are contained in the dataset published by the ECB as current account and savings account (including savings in building societies) information and therefore are used as the basis of comparison in this study. However, the ECB dataset does not contain any information about the allocation of households' savings to the individual sectors of the Austrian banking system, which is only available internally. The results of the comparison based on the second set of questions (amounts held at different banks) are in the Appendix B to this study as a sensitivity analysis and in general confirm the findings of the article.

As explained above, in the HFCS households were asked to indicate which banks they use rather than specifying the amounts held in individual accounts. If a household has several accounts at one and the same bank, the dataset records a customer relationship with a single bank. If a household has accounts at different banks, the dataset reflects customer relationships with several banks. The overwhelming majority of Austrian households use only a single bank – more than 91% of respondents in the HFCS – and only two percent of households have accounts with more than two different banks. However, households can be expected to have more than a single account with their so-called house bank. The first bank recorded, that is, the one at which the household holds the highest volume of funds, is also the one to which households are classified for the results in the article.

The deposit aggregates may be subdivided into sight accounts and savings deposits by bank sectors on the basis of the administrative account data that Austrian banks report to the OeNB. In addition, the total in savings accounts (only totals of domestic nonbanks, which include the self employed and sole proprietorships) may be further subdivided by deposit ranges. The data of the following bank sectors may be analysed separately:

- Joint stock banks
- Savings banks
- Raiffeisen credit cooperatives
- Volksbank credit cooperatives
- State mortgage banks
- Other

Raiffeisen and Volksbank are two types of credit cooperatives in the form of multistaged banks, which each form one separate banking sector in the banking statistics. Building societies are classified under the respective sector of the households' (house) bank, as customers associate their building society savings plans with their (house) bank. The category "other" is differently defined for the results from the HFCS and the banking statistics. In the HFCS, the households could choose to have a customer relationship with a bank from a predefined list of the 21 largest banks in Austria. In case the household wanted to state a different bank, a verbatim recording was available. If a respondent left the verbatim recording blank, the relationship was classified in the "other" category, since these responses could not be attributed to a banking sector ex post. In the banking statistics, "other" refers to special-purpose banks and banks as defined in Article 9 of the Austrian Banking Act (credit institutions from EU Member States). If a household has provided information about one of these banks in the verbatim text field, it was also classified to the category "other". Given the different definitions, no comparisons of this category were made; it is provided simply for the sake of completeness. Deposits can be allocated to the following ranges based on the administrative account data (the HFCS permits any type of classification):

- Up to €10,000
- €10,000 to €20,000
- €20,000 to €50,000
- €50,000 to €100,000
- €100,000 to €500,000
- €500,000 to €1,000,000
- €1,000,000 to €3,000,000
- Over €3 million

With data available in the banking statistics on both the number of accounts and the total volume of deposits, it is possible to calculate the average deposit holdings per account in a given deposit range for each and every bank sector separately. This average can be compared with the HFCS results for individual households. Due to the differences in the unit (account vs. household), however, one is expecting differences in the overall statistics since (potentially) several accounts are held by a single household (as explained above). Given the structure of the HFCS, where all accounts of a household are totalled, it might be expected that average deposits tend to be higher.

Appendix B: Additional Results

This appendix features three tables that repeat the calculations in Tables 4, 6, and 7 on the basis of the second way the information on the amounts (savings deposits) held at different banks was surveyed in the HFCS (see Appendix A). The use of data from this alternative survey method in the HFCS does not change the basic findings of the comparison of the HFCS and the financial accounts data. The appendix simply provides a sensitivity analysis for the classification of a household to a bank and for the different coverage methods of savings deposits.

Table A1. Aggregate savings deposits in individual banking sectors and deposit range

HFCS – total deposits, € million (variables for Austria)

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	67,799	6,179	5,830	12,037	9,619	19,103	8,308	n.a.[1]	.[2]
Joint stock banks	15,826	2,045	1,609	3,197	3,198	5,014	.	.	.
Savings banks	14,943	1,470	1,609	3,374	1,911	4,715	.	.	.
Raiffeisen credit cooperatives	22,440	1,986	1,839	3,831	2,705	4,023	n.a.	n.a.	.
Volksbank credit cooperatives	5,049	524	441	1,092	831	n.a.	.	.	.
State mortgage banks	1,527	168	164	n.a.	.	.	.	.	.
Other (national)	8,015	578	630	1,260	967	2,240	n.a.	.	.

Banking statistics – total deposits, € million

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	156,217	40,859	43,431	24,667	18,425	20,180	3,308	3,004	2,345
Joint stock banks	39,032	11,168	15,100	3,931	3,051	3,504	651	740	887
Savings banks	41,490	10,935	10,709	6,308	4,973	5,536	1,058	1,086	885
Raiffeisen credit cooperatives	56,118	13,744	12,379	11,682	7,997	8,194	1,057	745	319
Volksbank credit cooperatives	13,724	4,137	3,747	1,783	1,561	1,817	318	227	134
State mortgage banks	5,765	861	1,476	944	827	1,115	219	204	120
Other (national)	87	13	20	19	16	13	4	2	0

HFCS – deposits, share in % of Banking statistics figures

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	43.40%	15.12%	13.42%	48.80%	52.21%	94.66%	251.18%	.	.
Joint stock banks	40.55%	18.31%	10.66%	81.33%	104.82%	143.07%	.	.	.
Savings banks	36.02%	13.44%	15.02%	53.48%	38.43%	85.17%	.	.	.
Raiffeisen credit cooperatives	39.99%	14.45%	14.86%	32.79%	33.82%	49.10%	.	.	.
Volksbank credit cooperatives	36.79%	12.67%	11.77%	61.26%	53.22%	.	.	.	.
State mortgage banks	26.49%	19.50%	11.11%	.	.	.	.	.	.

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "n.a." have fewer than six observations; these values are not shown.

[2] All cells marked with "." have no observations.

Table A2. Allocation of customer relationships/accounts to banking sectors and deposit range

HFCS – share of accounts (variables for Austria)

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	100.0%	64.9%	13.1%	13.1%	4.7%	3.5%	0.4%	n.a.[1]	.[2]
Joint stock banks	40.4%	22.6%	3.6%	3.5%	1.5%	0.8%	.	.	.
Savings banks	31.2%	16.2%	3.6%	3.4%	0.8%	0.9%	.	.	.
Raiffeisen credit cooperatives	39.4%	20.3%	4.2%	4.2%	1.3%	0.8%	n.a.	n.a.	.
Volksbank credit cooperatives	10.0%	5.0%	0.9%	1.1%	0.4%	n.a.	.	.	.
State mortgage banks	2.6%	1.5%	0.4%	n.a.	.	.	.	.	.
Other (national)	14.2%	6.2%	1.4%	1.4%	0.5%	0.4%	0.1%	.	.

Banking statistics – share of accounts

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 to €3,000,000	Over €3,000,000
Total	100.0%	81.2%	13.7%	3.4%	1.2%	0.5%	0.0%	0.0%	0.0%
Joint stock banks	32.9%	27.2%	4.9%	0.6%	0.2%	0.1%	0.0%	0.0%	0.0%
Savings banks	26.1%	21.3%	3.4%	0.9%	0.3%	0.1%	0.0%	0.0%	0.0%
Raiffeisen credit cooperatives	30.5%	24.4%	3.8%	1.6%	0.5%	0.2%	0.0%	0.0%	0.0%
Volksbank credit cooperatives	8.3%	6.7%	1.1%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%
State mortgage banks	2.3%	1.7%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Other (national)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with "n.a." have fewer than six observations; these values are not shown.

[2] All cells marked with "." have no observations.

Table A3. Average deposit holdings by banking sectors and deposit range

HFCS – average deposit holdings (variables for Austria), €

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 €3,000,000	Over €3,000,000
Total	23,696	3,328	15,513	32,021	72,074	189,965	669,254	n.a.[1]	.[2]
(Std Err)	*4,252*	*147*	*323*	*793*	*2,429*	*16,553*	*65,978*	n.a.	.
Joint stock banks	13,687	3,168	15,747	31,665	73,795	220,952	.	.	.
(Std Err)	*2,276*	*205*	*562*	*1,374*	*4,783*	*31,335*	.	.	.
Savings banks	16,709	3,165	15,747	34,210	80,422	190,145	.	.	.
(Std Err)	*4,478*	*241*	*503*	*1,399*	*5,453*	*33,717*	.	.	.
Raiffeisen credit cooperatives	19,889	3,424	15,314	31,712	70,449	173,165	n.a.	n.a.	.
(Std Err)	*4,463*	*217*	*558*	*1,254*	*3,747*	*27,152*	n.a.	n.a.	.
Volksbank credit cooperatives	17,693	3,650	16,386	35,654	69,600	n.a.	.	.	.
(Std Err)	*7,986*	*468*	*1,006*	*3,068*	*7,513*	n.a.	.	.	.
State mortgage banks	21,049	3,986	14,329	n.a.	.	.	.	.	.
(Std Err)	*21,541*	*844*	*1,384*	n.a.	.	.	.	.	.
Other (national)	19,822	3,241	15,660	31,757	66,150	180,061	n.a.	.	.
(Std Err)	*7,028*	*393*	*717*	*1,698*	*4,991*	*38,976*	n.a.	.	.

Banking statistics – average deposit holdings, €

	Total	Up to €10,000	€10,000 to €20,000	€20,001 to €50,000	€50,001 to €100,000	€100,001 to €500,000	€500,001 to €1,000,000	€1,000,001 €3,000,000	Over €3,000,000
Total	6,658	2,144	13,539	30,909	67,867	168,288	659,002	1,530,099	6,319,995
Joint stock banks	5,057	1,753	13,126	30,305	68,117	166,951	663,942	1,564,715	7,709,296
Savings banks	6,787	2,186	13,411	31,058	66,302	172,241	664,159	1,565,499	6,319,543
Raiffeisen credit cooperatives	7,852	2,405	14,029	30,726	67,828	163,988	645,182	1,440,230	4,906,600
Volksbank credit cooperatives	7,078	2,616	13,901	32,504	71,089	172,677	664,868	1,445,637	5,601,417
State mortgage banks	10,644	2,219	13,977	31,921	71,279	179,325	677,077	1,684,248	4,446,296
Other (national)	17,556	4,944	13,805	31,903	70,347	184,056	848,000	1,660,000	.

Source: HFCS Austria 2010, OeNB; OeNB banking statistics.

Note: Savings plans with building and loan associations are aggregated to the appropriate sectors.

[1] All cells marked with “n.a.” have fewer than six observations; these values are not shown.

[2] All cells marked with “.” have no observations.

7. References

Albacete, N., P. Lindner, K. Wagner, and S. Zottel. 2012. “Eurosystem Finance and Consumption Survey 2010: Methodological Notes for Austria.” *Addendum to Monetary Policy and the Economy* Q3/12: 1–100.

Andreasch, M., P. Fessler, and M. Schürz. 2009. “Austrian Households’ Equity Capital – Evidence from Microdata.” *Monetary Policy and the Economy* Q4/09: 61–78.

Antoniewicz, R., R. Bonci, A. Generale, G. Marchese, A. Neri, K. Maser, and P. O’Hagan. 2005. “Household Wealth: Comparing Micro and Macro Data in Canada, Italy and United States.” Paper prepared for the LWS Workshop: “Construction and Usage of Comparable Microdata on Wealth: the LWS”, Banca d’Italia, Perugia, Italy, 27–29 January 2005. Available at: http://www.lisproject.org/lws/introduction/files/antoniewiczrevised.pdf (accessed 6 November 2012).

Avery, R.B., G.E. Elliehausen, and A.B. Kennickell. 1988. “Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finance.” *Review of Income and Wealth* 34: 339–369.

Bricker, J. and G.V. Engelhardt. 2008. “Measurement Error in Earnings Data in the Health and Retirement Study.” *Journal of Economic and Social Measurement* 33: 39–61.

Davies, P.S. and L.T. Fisher. 2009. “Measurement Issues Associated with Using Survey Data Matched with Administrative Data from the Social Security Administration.” *Social Security Bulletin* 69: 1–12.

ECB. 2013. “The Eurosystem Household Finance and Consumption Survey – Results from the First Wave.” *Statistics Paper Series ECB* 2: 1–111.

Fessler, P., M. Kasy, and P. Lindner. 2012. “Survey mode effects on income inequality measurement.” Paper prepared for the 32nd General Conference of The International Association for Research in Income and Wealth. Available at: http://www.iariw.org/papers/2012/LindnerPaper.pdf (accessed 2 December 2013).

Fessler, P., P. Mooslechner, and M. Schürz. 2012. “Eurosystem Household Finance and Consumption Survey 2010 – First Results for Austria.” *Monetary Policy and the Economy* Q3/12: 24–62.

Henriques, A.M. and J.W. Hsu. 2014. “Analysis of Wealth Using Micro- and Macrodata: A Comparison of the Survey of Consumer Finances and Flow of Funds Accounts.” In *Measuring Economic Sustainability and Progress*. National Bureau of Economic Research Studies in Income and Wealth (Book 72), edited by D.W. Jorgenson, J. Steven Landefeld, and P. Schreyer, 245–276. Chicago and London: University of Chicago Press.

Johansson, F. and A. Klevmarken. 2007. “Comparing Register and Survey Data.” Chapter 2 in *Essays on Measurement Errors and Nonresponse*. Economic Studies 103. Ph.D. diss., Department of Economics, Uppsala University.

Kapteyn, A. and J.Y. Ypma. 2007. “Measurement Error and Misclassification: A Comparison of Survey and Administrative Data.” *Journal of Labor Economics* 25: 513–551.

Kavonius, I.K. and J. Honkkila. 2013. “Reconciling Micro and Macro Data on Household Wealth: A Test Based on Three Euro Area Countries.” *Journal of Economic and Social Policy* 15: 1–30.

Kavonius, I.K. and V-M. Törmälehto. 2010. "Integrating Micro and Macro Accounts – The Linkages between Euro Area Household Wealth Survey and Aggregate Balance Sheets for Households." Prepared for the 31st General Conference of the International Association for Research on Income and Wealth, St. Gallen, 22–28 August 2010. Available at: http://www.iariw.org/papers/2010/7aKavonius.pdf (accessed 16 April 2015).

Kavonius, I.K., and V-M. Törmälehto. 2003. "Household Income Aggregates in Micro and Macro Statistics." *Statistical Journal of the United Nations Economic Commission for Europe* 20: 9–25.

Mathä, T.Y., A. Porpiglia, and M. Ziegelmeyer. 2012. "The Luxembourg Household Finance and Consumption Survey (LU-HFCS): Introduction and Results." Cahier D'Études Working Paper, Banque Centrale Du Luxembourg.

Roemer, M. 2002. "Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of the Income and Program Participation." *U.S. Census Bureau Technical Paper* No. TP-2002-22. Available at: https://www.census.gov/hhes/www/income/publications/asa2002.pdf (accessed 16 April 2015).

Sierminska, E., A. Brandolini, and T.M. Smeeding. 2006. "Comparing Wealth Distribution across Rich Countries: First Results from the Luxembourg Wealth Study." Luxembourg Wealth Study Working Paper Series 1. Available at: http://www.lisdatacenter.org/wps/lwswps/1.pdf (accessed 16 April 2015).

Stiglitz, J.E., A. Sen, and J.-P. Fitoussi. 2009. "Report by the Commission on the Measurement of Economic Performance and Social Progress." CMEPSP. Available at: http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf (accessed 16 April 2015).

Törmälehto, V.-M. 2011. "Comparing National Accounts Income Aggregates with Income Aggregates based on LIS Micro-Data." LIS Technical Working Paper Series 2. Revised version.

Received July 2014
Revised May 2015
Accepted August 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 29–73, http://dx.doi.org/10.1515/JOS-2016-0002

Respondent-Driven Sampling – Testing Assumptions: Sampling with Replacement

Vladimir D. Barash[1], *Christopher J. Cameron*[2], *Michael W. Spiller*[3], *and Douglas D. Heckathorn*[4]

Classical Respondent-Driven Sampling (RDS) estimators are based on a Markov Process model in which sampling occurs with replacement. Given that respondents generally cannot be interviewed more than once, this assumption is counterfactual. We join recent work by Gile and Handcock in exploring the implications of the sampling-with-replacement assumption for bias of RDS estimators. We differ from previous studies in examining a wider range of sampling fractions and in using not only simulations but also formal proofs. One key finding is that RDS estimates are surprisingly stable even in the presence of substantial sampling fractions. Our analyses show that the sampling-with-replacement assumption is a minor contributor to bias for sampling fractions under 40%, and bias is negligible for the 20% or smaller sampling fractions typical of field applications of RDS.

Key words: Respondent-driven sampling; hidden populations; sampling with replacement.

1. Introduction

Respondent-Driven Sampling (RDS) has become the method of choice for studies of hidden and hard-to-reach populations, yet important questions regarding the method remain unresolved. RDS is a form of network sampling paired with an estimation strategy, where individuals are treated as network nodes and their social relationships are treated as edges.

Drawing an RDS sample involves several steps. First, when sampling from a hidden population, one begins with a convenience sample of initial respondents who serve as “seeds”. Seeds can be identified by key informants who are drawn from organizations where the target population congregates, or they may self-identify by volunteering for the study. Second, initial respondents each recruit several peers, who compose the sample’s first “wave”. Third, the first-wave recruits each recruit several peers, who form the sample’s second wave. Thus the first-wave recruits become the recruiters of the second wave. Fourth, the sample expands in this recursive manner, wave by wave, with the prior wave’s recruits becoming the recruiters of the subsequent wave, until the desired sample size has been reached.

[1] Graphika Inc., 116 West 23rd Street, 5th Floor, New York NY 10011, U.S.A. Email: vlad.barash@graphika.com
[2] Cornell University – Sociology, 344 Uris Hall Ithaca, New York 14853, U.S.A. Email: cjc73@cornell.edu
[3] Cornell University – Sociology, 344 Uris Hall, Ithaca, New York 14853, U.S.A. Email: mws24@cornell.edu
[4] (Corresponding author) Cornell University – Sociology, 344 Uris Hall Ithaca, New York 14853, U.S.A. Email: douglas.heckathorn@cornell.edu

Acknowledgments: This research was made possible by a grant from the National Institutes of Health/National Institute for Nursing Research (1R21NR10961). We thank members of the Symposium on Respondent-Driven Sampling, Department of Mathematics, Stockholm University for helpful comments and advice.

One essential feature of the sampling method includes keeping track of who has recruited whom. This is important because affiliation patterns (e.g., members of a racial/ethnic group tending to recruit members of the same racial/ethnic group) affect the composition of the sample. A second essential feature of the sampling method is asking each respondent how many members of the target population they know as acquaintances, friends, or closer than friends. The people in the target population known to an individual node defines the node's network neighborhood, and the number of nodes in the neighborhood defines the node's degree. Nodes of larger degree tend to be oversampled because they have a larger number of edges that serve as peer-recruitment paths.

The RDS estimators employ information about a respondent's degree and their affiliation patterns to correct for sources of bias inherent in chain-referral sampling. Specifically, in the computation of the RDS estimator, the estimated size of each of the population's subgroups is inflated or deflated based on whether the subgroup was judged to be under- or oversampled. In sum, the RDS estimator functions some what like a corrective lens that compensates for network-based sources of bias in the sampling process.

The advantage of RDS is that it provides a means for drawing probability samples of populations which cannot be effectively sampled using traditional population survey methods because they lack a sampling frame, and because these populations have social networks that are hard for outsiders to penetrate due to stigma or privacy concerns.

RDS studies have focused both on populations of relevance to public health, such as injection drug users (IDUs), men who have sex with men (MSM), and commercial sex workers (CSW); on populations of relevance to arts and culture such as jazz musicians (Heckathorn and Jeffri 2001) and visual artists (Jeffri et al. 2011); on hard-to-reach or rare general populations such as low-wage workers (Bernhardt et al. 2012) and Canadian urban aboriginals (Smylie et al. 2011); and on criminological populations such as underage sex trafficking victims (Curtis et al. 2008). A 2009 survey (Malekinejad et al. 2008) analyzed the results of 128 studies drawn from more than 28 countries. RDS has been employed in studies funded by agencies including, the Centers for Disease Control and Prevention (CDC), CDC/Global AIDS, Gates India, the United States Agency for International Development (USAID), the National Science Foundation (NSF), and National Institutes of Health (NIH) institutes including the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH), the National Institute on Child Health and Human Development (NICHD) and the National Institute of Nursing Research (NINR).

The popularity of RDS derives in large part from a proof (Salganik and Heckathorn 2004) showing that when the assumptions of the method are satisfied, population estimates are asymptotically unbiased. This means that bias is only on the order of $1/n$, where n is the sample size, so bias is trivial in samples of significant size. A subsequent paper (Heckathorn 2007) reduced by one the number of assumptions required by the method, so the proof of lack of bias depends on four conditions: (1) the network connecting the population is dense enough to form a single component; (2) recruiters know one another, as acquaintances, friends, or those closer than friends, so their relationships are reciprocal; (3) respondents recruit as though they are selecting randomly from their neighborhoods; and (4) sampling occurs with replacement. In this article, unless otherwise specified, we use the Heckathorn 2007 estimator because it requires fewer assumptions than other estimators, and it controls for a form of bias ignored by other RDS estimators, differential recruitment by degree. (For a comparison of estimators, see Heckathorn 2011.)

The first three assumptions can be approximated given a suitable choice of population to study and of research design. Given its reliance on network-based recruitment, RDS is a method suited only for studying populations with relatively dense networks. Generally, this involves populations united by a contact pattern, that is, network connections created by virtue of membership in the population. For example, drug users form ties when purchasing and using drugs and jazz musicians form ties when performing in an ensemble. When membership in a population does not create contact patterns, as is the case for tax evaders, the population is not suitable for study using RDS or any other network-based method. Hence, assumption one determines the conditions under which RDS is a suitable sampling methodology.

The second assumption, that respondents know one another well enough for their relationships to be reciprocal, can be satisfied by appropriate research design. This involves making recruitment rights both scarce and valuable. Generally this is done through a combination of rewards for peer recruitment and quotas to limit the number of peers who can be recruited. Quotas are implemented by giving each respondent, that is, each potential recruiter, a limited number of recruitment coupons. Each coupon has a unique serial number which allows the recruiter to whom the coupon was given to be linked to the recruit who brings it into the study site. When recruitment rights are scarce and valuable, few respondents waste them on strangers who may fail to take advantage of the recruitment opportunity. Consequently, recruitment by strangers is generally infrequent—less than a few percent (Iguchi et al. 2009)—and these recruitments can be identified by asking respondents about their relationship to their recruiter, and then deleting these cases to produce a data set consisting exclusively of respondents who know one another, as acquaintances, friends, or closer than friends. The rationale for the reciprocity assumption is that for any individual, those who consider him or her as acquaintances or friends (i.e., the individual's "in-degree"), also tend to be viewed by the individual as an acquaintance or friend (i.e., the individual's "out-degree"). Because in-degree and out-degree are equivalent in networks where all ties are reciprocal, we refer merely to "degree".

The third assumption, that respondents recruit randomly from their neighborhoods, can be best approximated when opportunities to be interviewed are easily and safely available to all members of the neighborhood because respondents have no incentive to selectively favor or exclude any particular neighbors (Heckathorn 2007, 163–164). Here it is important to note that individuals are not assumed to recruit randomly from the target population, for this target population is generally far larger than any node's neighborhood. Furthermore, the composition of nodes' neighborhoods varies greatly, because those similar in race/ethnicity, education, income, religion, and other factors tend to affiliate, so neighborhoods are often relatively homogeneous, a factor termed homophily. Hence, when a node recruits randomly from its neighborhood, this does not mean that it is sampling randomly from the target population. Support for this assumption of random recruitment from neighborhoods has been found in several studies (e.g., Heckathorn et al. 2002; Wang et al. 2005); however, the conditions under which it holds or is violated warrant further study.

The final assumption, that sampling occurs with replacement, has a unique status because it is invariably counterfactual irrespective of choice of population or research design; respondents can only be interviewed once, and hence replacement is excluded. Sampling with replacement is not feasible in practice because RDS survey respondents are generally compensated for their participation and allowing a single individual to participate multiple times could lead to strategic recruiting behavior by participants. Methods have been developed to reduce subject duplication through a database that records scars, tattoos, and biomarkers (Heckathorn et al. 2001). These methods are necessary in RDS sampling because most RDS studies require subject anonymity due to the sensitive nature of questions in these studies: some RDS studies have asked about sexual and drug use history (Iguchi et al. 2009), while others have asked employees to report their employers' violations of workplace laws (Bernard et al. 2010). Even if sampling with replacement did not create perverse recruitment incentives, it would be inefficient and costly to interview a person each time they were recruited. Compared to a hypothetical implementation of with-replacement RDS, without-replacement RDS yields data from a greater number of unique individuals.

If respondent-driven sampling were conducted with replacement and in accordance with the other assumptions, the RDS process would be a simple random walk on the network conforming to a Markov Process. Previous studies (Volz and Heckathorn 2008) have assumed that, for very small sampling fractions, the sampling-with-replacement assumption is valid; in our analysis, we put these statements to the test and, more broadly, explore the potential for bias resulting from this counterfactual assumption.

The general expectation in statistical analyses is that a sampling-with-replacement assumption, which is equivalent to the assumption of an infinite population size (Gile 2011, 32), becomes more problematic as the sampling fraction increases. This may suggest that the replacement assumption does not significantly bias RDS studies with small sampling fractions. For example, in the CDC's National HIV Behavioral Surveillance Injection Drug User (NHBS-IDU) study, the sampling fraction for the 23 study sites had a median of 2.3% and a range of 0.6% to 8% (Lansky et al. 2009). Similarly, in the NEA-funded study of jazz musicians, the sampling fractions were 0.8% and 1.6% in New York and San Francisco, respectively (Heckathorn and Jeffri 2001). However, given the unique nature of RDS sampling, especially the interdependence of observations owing to the tendency of respondents to recruit others like themselves, there can be no guarantee that this rule of thumb is valid.

Furthermore, there are inevitably studies in which the research design calls for a large sample from a small population. In such cases, the sampling fraction is large. The largest sampling fraction with which this research team has been directly involved was a study of IDUs in a small Connecticut town with a population of 59,000 of whom slightly more than 1,000 were injectors. The sampling fraction was estimated at a rather substantial 37% via capture-recapture using a combination of RDS and police statistics (Heckathorn et al. 2002).

In this article, we extend work by Gile (2011) and Gile and Handcock (2010) in analyzing bias introduced by violation of the sampling-with-replacement assumption, as well as overall bias of RDS estimates. We employ formal proofs complemented with simulation to explore the relationship between bias and sampling fraction, including factors that affect this relationship, such as network density, homophily and the degree distribution in the population. Our results suggest that under certain circumstances, RDS estimates are surprisingly stable even in the presence of high (i.e., 50% or greater) sampling fractions.

Previous work, such as the Successive Sampling Estimator (Gile 2011), has explored solutions to reducing the bias of the Volz-Heckathorn (VH) estimator (Volz and Heckathorn 2008) for high sampling fractions. Such work is a useful exploration of the boundary conditions for RDS studies; however, a major limitation is that it focuses exclusively on sampling fractions of 50% to 95%, and does not explore the magnitude of bias in low and moderate sampling fractions. Sampling fractions approaching 100% are more akin to census-level studies, where the population proportions can be calculated directly, than to random samples requiring population estimators. In the limit of 100% sampling fraction (census), a simple sample proportion gives the correct prevalence value without needing an estimator. Therefore, while the Successive Sampling Estimator performs well at fractions over 50%, it is important to note that for some of the range of sampling fractions the Gile 2011 study covers, census-level measures, such as simple sample proportion, are often more appropriate for parameter estimation than RDS estimators such as VH or successive sampling. A variance-estimation method appropriate for RDS is still required as simple random sample variance estimates will be much too small. Note that Gile (2011) compares the Successive Sampling Estimator to the VH estimator. As long as all groups recruit equally effectively (as is the case in Gile's simulations and in this article's simulation), the VH and Heckathorn (2007) estimators produce identical point estimates and may be directly compared (Volz and Heckathorn 2008; Heckathorn 2007). In the discussion section of this study below, we address in more detail the differences between the Successive Sampling Estimator, the Heckathorn (2007) estimator and the simple sample proportion.

What we explore in this article is a range of sampling fractions not examined in previous studies, which were limited to a range from 50% to 95% (Gile 2011). Our key finding is that RDS estimates within the parameter ranges we examine (i.e., sampling fractions between 5% and 80%) are surprisingly unbiased even in samples with high (i.e., 50% or greater) sampling fractions. We did not examine sampling fractions in excess of 80%, because few field studies reach such a high proportion of the population. Gile (2011) has already examined this upper range and demonstrated that very large sampling fractions result in significant VH estimator bias. We found that expected bias is less than five percentage points away from the true population proportion across the ranges of sampling fraction, homophily, and degree distribution examined in our simulations.

We also find that violation of the sampling-with-replacement assumption is not a major contributing factor to bias in RDS simulations for sampling fractions under 40%; other factors, such as relative mean degree of the target group, tend to affect bias more than sampling without replacement in these conditions. In particular, both the mean and the 95% confidence intervals for sampling with and without replacement are essentially identical (within 1–2% of each other) for sampling fractions up to 20%.

Finally, we analyze the sampling variability of the RDS method, which we define as the width of the 95% confidence interval of the sampling distribution over a set of simulated RDS samples, using the Heckathorn (2007) estimator to calculate sample prevalence. We find that the sampling variability of RDS decreases with increasing sampling fraction for sampling fractions under 40%. In combination, these findings suggest an optimal range of one to 20% of sampling fractions for RDS studies using the Heckathorn (2007) estimator.

The rest of this article is organized as follows: We first present a formal model of an RDS sampling process, with a few additional assumptions for simplicity. We use this model to derive several results about the extent of bias due to sampling without replacement in different network structures, for different values of sampling fraction and (for a two-group system) homophily and relative group size. Next, we confirm our formal results with simulations, and extend them further by offering simulations that go beyond the scope of our proofs. Finally, we offer a brief discussion of the analysis with implications for future RDS studies in the field, and conclude with directions for future work in this area.

2. Formal Model

For reference, we begin this section with a glossary of terms and notation used throughout the rest of the article. Notation is presented in Table 1.

Network Terms:

Social Network. The social network is the set of individuals in a population (nodes) and a set of connections (friendships, acquaintances, etc.) between these individuals (edges).

Node. In a social network, a person corresponds to a particular node in the network.

Edge. Edges connect nodes in a social network and represent the relationships between the individuals represented by the nodes. Relationships (i.e., friendship) can be undirected, such that an edge from node A to node B implies an edge from B to A, or directed. The social networks we consider are assumed to have undirected edges.

Path. Sequence of consecutive edges connecting two nodes in a network. For example, if we have a network of three individuals A,B,C with edges between A and B and B and C, then a path exists between A and C.

Neighborhood/Neighbors. For any node *ego* in a network, the nodes directly connected to ego via an edge define ego's neighborhood. Nodes connected by an edge are considered neighbors.

Degree. For any node *ego* in a network, ego's degree is the number of nodes in ego's neighborhood.

Degree Distribution. For a collection of nodes in the graph, usually the entire population or a sample, the degree distribution is a probability distribution $P(k)$ where $P(k)$ is the fraction of nodes in the collection with degree equal to k. Degree distributions are necessarily discrete, but it is common to approximate the shape of $P(k)$ with a continuous probability distribution. Common models of degree distribution include the uniform distribution, in which every node has the same degree, the power-law degree distribution, and the Poisson degree distribution.

Table 1. Notation

Symbol	Meaning
r	An individual recruit in an RDS sample, or participant in RDS study
r_i	The *i*th recruit in an RDS sample, or participant in RDS study, in order of recruitment
R_j	*j*th wave of recruiters for an RDS study. If $j = 0$, this is the set of "seed nodes" for the RDS study
R	The full collection of recruiters and recruits in an RDS sample, ordered by recruitment
NR	The number of recruiters in an RDS study
P	The larger population from which the RDS sample is drawn
N	The total number of potential recruits in an RDS study
G	The graph of potential recruits in an RDS study
$d(i)$	Degree of node i (representing recruit r_i) in G
$\mu_d(A)$	Mean degree of a set A of nodes (recruits) in G
L	The node-level variable of interest in an RDS study, for example HIV Status
$L(r)$	Value of L for recruit r, for example, "HIV-positive"
S	The set of all distinct values of L present in R, for example, {"HIV-positive", "HIV-negative", "unknown"}
s_k	s is an enumeration of S and k is an index variable for the elements of s. $k \in \{1,\ldots,\lvert S\rvert\}$
$p(s_c, s_d)$	Transition probability between two states of the Markov Process model of RDS, where individual states correspond to elements of S
$\pi(s_k)$	The proportion of recruits in R with value of $L = s_{k,}$ a.k.a. the sample proportion of recruits with value $L = s_k$
σ	The proportion of the population in the sample. $\sigma = \lvert\{r\colon r \in R\}\rvert/N$
$\lfloor\sigma N\rfloor$	Sample size for a particular RDS sample
ω	Repeated sampling event where participant r_i in an RDS study attempts to recruit participant r_j but discovers that r_j has already been recruited
ρ_ω	Density of repeated sampling events in a particular RDS study or simulation

Poisson Degree Distribution. A degree distribution where the fraction of nodes with degree k follows the Poisson probability distribution, that is $(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, where λ is equal to the mean node degree.

Power-Law Degree Distribution. A degree distribution where the fraction of nodes with a degree k follows a Power Law; $(k) \sim k^{-\gamma}$, where γ is a shape parameter.

Homophily (group-level attribute). The homophily index is the mathematical value capturing the extent to which individuals in a particular group are connected to nodes within the group rather than nodes in other groups. A homophily value of zero means the proportion of ties between members of a group is consistent with the proportion of within-group ties that would be expected if ties were formed at random. A positive homophily value indicates the presence of an in-group affiliation bias and a value of 1 indicates that the group is entirely isolated from other groups—all ties from the group are to other members of the group.

Tree. Named for the characteristic branching shape, a tree is a type of network where any two nodes are connected by exactly one path (ignoring the directionality of the edges if the edges are directed).

RDS Terms:

Recall that respondent-driven sampling starts with a set of seed nodes selected by some nonrandom process. Each seed node selects k random nodes from its neighborhood to generate a collection of new participants in a process called *recruitment*. In recruitment, the nodes selecting new participants are referred to as *recruiters* and the selected nodes are referred to as *recruits*. The seeds constitute sample wave 0 and the collection of nodes selected by the seeds constitutes sample wave 1. The nodes in wave 1 recruit wave 2 and so on with each new wave of recruits serving as recruiters for the next wave. Except for the seeds and the final wave of recruits (who do not recruit in turn), each participant serves as both a recruiter and as a recruit.

Seed. A member of the initial wave of recruiters in an RDS sample.

Recruit. A member of the second through last wave of recruiters in an RDS sample.

Recruitment Event. RDS is a type of chain-referral sample where individuals already in the sample attempt to recruit new participants from their own neighborhoods. We call each of these recruitment attempts a recruitment event. The recruitment event may succeed or fail. A node in the sample generates a recruitment event each time it tries to recruit a neighbor, so a pair of nodes may generate duplicate recruitment events when sampling with replacement. A successful recruitment by a node in sample wave w results in a new member in sample wave $w + 1$. Respondent-driven sampling only captures successful recruitment events.

Sample Chain. A sequence of edges and nodes that reflects a series of recruitments. Every recruit in an RDS sample can be traced back through a series of recruiter/recruit relationships to a seed node. When sampling is without replacement, the chain is unique and the set of nodes and edges from all the chains that begin from a particular seed define a directed tree graph with the seed as the root node.

Group. The set of individuals with the same value for a particular categorical variable (L). Membership in a group does not imply relationships to other members of the same group and ties between nodes do not imply membership in the same group. The population can be partitioned into groups based on a variable of interest (L) with the number of groups equal to the number of unique levels of (L). If the variable of interest were race, then the groups would correspond to unique race groups. If the variable of interest were HIV status, then the groups might correspond to “Positive”, “Negative” and “Don’t Know.” The point of RDS analysis is to produce an estimate of the relative size of each group in the population.

Recruitment relationship. With the exception of the sample seeds, each individual recruited into the sample is recruited by someone else already in the sample. Each successful recruitment event generates a recruitment relationship in the sample between the recruiter and the recruit. For a given individual-level categorical attribute (L), the recruitment relationships in the sample can be labeled by the value of L for the recruiter and the recruit. Recall that each individual in the population has an attribute corresponding to a value for categorical variable L so each sample member $r \in R$ can be labeled by L(r). For any particular categorical variable L, recruitment events in the sample can be labeled by the tuple (L(recruiter), L(recruit)). When L is a binary variable with levels 0 and 1, then there are four possible labels for a recruiter → recruit relationship: (0,0), (0,1), (1,0), (1,1).

Markov Process. A stochastic process whose next state depends only on its present state. RDS recruitment is a Markov Process because the future state of the process (future recruits) is dependent only on the current state (current recruits) via the recruitment mechanism.

Irreducible Markov Process. A Markov Process for which it is possible to get to any state from any state. RDS Recruitment is not an irreducible Markov Process because individuals cannot be recruited more than once, so once the process has gotten to a particular state (a particular set of recruits) it cannot return to that state.

Notation:

We use $|X|$ to indicate the number of elements in collection X. For collections that may contain duplicates, the size of the collection is the number of elements in the collection including duplicate entries.

We use set builder notation to describe sets, as in $S = \{L(r)\colon r \in R\}$, by which we mean S is the set of unique values produced by applying the function $L(r)$ to each element in the collection R. We also need to construct collections that may contain duplicate elements. For these collections, we adopt an angle bracket notation as in $R_{\text{values}} = \langle L(r)\colon r \in R\rangle$ by which we mean the collection R_{values} is the sequence $(L(r_i))_{i=1}^{|R|}$ where r is an enumeration of the collection R in order of recruitment. Though S and R_{values} have the same number of distinct values, the number of elements in R_{values} is equal to the number of elements in R, while the number of elements in S is the number of distinct values of the categorical variable L. Stated more simply, S is a collection of the unique values in R_{values}.

The formal foundation for RDS as a Markov Process has been well explored in previous articles, such as Heckathorn (1997) and Volz and Heckathorn (2008). Previous work (Heckathorn 1997; Volz and Heckathorn 2008) showed that RDS can be modeled as a Markov Process. We do not repeat the analysis here for space reasons, but summarize it.

Assume that we wish to estimate the population composition in terms of some individual-level categorical variable L. The first step is to gather an RDS sample. Consider an arbitrarily selected seed set R_0 of initial recruiters for an RDS study. These recruiters are embedded in a larger population P, members of which are connected in a social network. For simplicity, we posit that recruitment into the RDS study can occur through any network connection—that is, if two individuals are network neighbors, it is possible for one to recruit the other. It is possible to define possible recruitment paths differently, but the choice of definition does not affect our formal analysis, so we use this simple one. Each member of collection R_0 then recruits a random subcollection of its neighbors.

The elements of these subcollections form the collection R_1 of recruits (which is also a subcollection of the network neighbors of R_0). The attributes of the individuals in R_1 are measured, including node degree, the value for variable L. The process then repeats, with R_1 recruiting a subcollection of the neighbors of R_1 to generate the collection R_2 and R_j generating the collection R_{j+1} until the RDS sampling process is stopped. At the end of the study, the full collection of recruiters and recruits R is an object of analysis.

The RDS sample is a collection of nodes R, the measured attributes of nodes in R and the recruitment relationships between recruiter and recruit. In order to estimate the transition probabilities for the underlying Markov Process, we need to construct a list of all the successful recruitment events captured by the sample. For each sample wave $j > 0$ each recruit $r_i \in R_j$ can be paired with at least one recruiter $r_h \in R_{j-1}$ who recruited r_i into the RDS sample, generating a recruitment event. When sampling without replacement, there is exactly one recruiter for each recruit. When sampling with replacement, a particular individual may appear multiple times within a single wave and may also appear in later waves. This is not a problem because there is exactly one successful recruitment event for each appearance of an individual in the sample. The collection of recruitment pairings between R_j and R_{j-1} for all waves j where $j > 0$ constitutes the list of recruitment events in the sample.

Each individual in the population has some value for the categorical attribute L so each sample member $r \in R$ can be labeled by L(r). The recruitment events in the sample can also be labeled by the sequence $\langle L(\text{recruiter}), L(\text{recruit})\rangle$.

Consider an RDS sampling process that starts from a single seed and where each recruiter generates at most one recruitment event. In this case, each sample wave $j{:}j > 0$ is generated by one recruitment event; R_j has exactly one element r_i and for any $r \in R_j{:}j > 0$, we can observe the node label of the ith recruit $L(r_i)$ and the label of the kth recruitment event $\langle L(r_{k-1}), L(r_k)\rangle$. The sequence of node labels produced by the sampling process can be modeled as a Markov Process on the distinct values of L.

Let S be the set of distinct values of L so $S = \{L(r){:}\ r \in R\}$. We can model the recruitment process as a Markov Process, where the states $s_k \in S$, $k = 1, 2 \ldots |S|$ correspond to distinct values of L and the transition from states s_c to s_d represent a recruitment event with label $\langle s_c, s_d\rangle$. Thus the underlying Markov Process that models *RDS* is between different values of an individual variable L, not between individual nodes. The work in Volz and Heckathorn (2008) models RDS as a Markov Process between individual nodes; however, the result of their analysis is identical to the Heckathorn (1997) analysis: that is, so long as the underlying Markov Process is irreducible, a stationary equilibrium exists where the state of the Markov Process (the current node) is independent of the starting state (the seed node). At this point, the steady state distribution of the Markov Process modeling the RDS recruitment process is an unbiased estimator of the population. Similarly, Volz and Heckathorn's analysis requires the assumption of sampling with replacement to be met for the irreducibility condition to be satisfied. Accordingly, the analysis in the rest of the article would be substantively the same were we to choose Volz and Heckathorn's (2008) model instead of the Heckathorn (1997) model. We focus on the Heckathorn (1997) model because it is much more straightforward to formulate bias due to sampling without replacement in terms of transitions between groups than it is to formulate the same in terms of transitions between nodes. More discussion follows in the Appendix.

We can also conceive of this process as occurring between groups *A, B,*. . . where an individual group contains all recruits in *R* who have the variable value s_c. In either case, previous work on RDS shows that as long as the Markov Process is irreducible, a condition that holds if a number of assumptions, including sampling with replacement, are satisfied, it will reach equilibrium. After the Markov Process reaches equilibrium, the state of the system is independent of the starting state, and recruits sampled after equilibrium will be independent of the seeds. A Markov chain is irreducible if it is possible to move from every state to every other state in a finite number of steps—that is, there can be no groups or sets of groups with homophily of 1. When these conditions hold, the mix of recruits will be independent of the seeds. In this case, if all individuals have equal degree, the RDS sample is representative of the underlying population.

The Markov Process model of RDS is also critical for estimating population composition by the levels of *L*. Heckathorn (1997) shows that it is possible to use the transition probabilities $p(s_c, s_d)$ between states in the Markov Process to construct a system of simultaneous equations that will yield the sample proportions $\pi(s_c)$ of recruits with key variable value equal to s_c. Furthermore, as the same work showed, one can estimate the transition probabilities $p(s_c, s_d)$ using the frequency of individuals in the collection $\langle r\text{: } L(r) = s_c\rangle$ recruiting individuals in the collection $\langle r\text{: } L(r) = s_d\rangle$ as follows:

$$\hat{p}_{RDS}(s_c, s_d) = \frac{\left|Rec\left(\langle r\text{: } L(r) = s_c\rangle, \langle r\text{: } L(r) = s_d\rangle\right)\right|}{\left|Rec\left(\langle r\text{: } L(r) = s_c\rangle, R\right)\right|} \qquad (1)$$

where $Rec(A,B)$ is the collection of all recruitment events where an individual in *A* recruits another individual in *B*. Later, Salganik and Heckathorn (2004) showed that if sampling occurs with replacement (any individual can be recruited any number of times), the RDS estimates of the Markov Process transition probabilities are unbiased, so as sample size increases, $\hat{p}_{RDS}(s_c, s_d)$ approaches the equilibrium transition probabilities $p_{MC}(s_c, s_d)$ of the underlying Markov Process. When R is the result of a Markov Process, $\hat{p}_{RDS}(s_c, s_d)$ is exactly the same as the maximum-likelihood estimate for the transition probabilities $\hat{p}_{MC}(s_c, s_d)$.

$$\hat{p}_{MC}(s_c, s_d) = \frac{\left|D\left(\langle r\text{: } L(r) = s_c\rangle, \langle r\text{: } L(r) = s_d\rangle\right)\right|}{\left|D\left(\langle r\text{: } L(r) = s_c\rangle, R\right)\right|} \qquad (2)$$

where $D(A,B)$ is the set of network connections between individuals in collection *A* and individuals in collection *B*.

However, if sampling occurs without replacement, the Markov Process model of RDS must be called into question. Since every individual may be recruited at most once, and there are a finite number of individuals, the underlying Markov Process is no longer irreducible, and thus does not have a stationary equilibrium distribution.

We cannot use the reducible Markov Process for sampling without replacement to calculate transition probabilities $p(s_c, s_d)$ and sample proportions $\pi(s_c)$. However, we can still use the irreducible Markov Process for sampling with replacement, and calculate transition probabilities and sample proportions for that process, as long as RDS chains are sufficiently similar to those that would be produced under sampling with replacement. To the extent that there is a difference between actual RDS chains and chain-referral samples with replacement, $\hat{p}_{RDS}(s_c,\ s_d)$ will not be an unbiased estimate of the true transition probabilities $p(s_c,\ s_d)$.

We note that $\hat{p}_{MC}(s_c,\ s_d)$ is the transition probability that any (not a particular) individual with key variable value s_c recruits any other (not a particular) individual with key variable value s_d. Therefore, it is a measure of transitions between groups of individuals and depends on the number of edges between these groups—in this sense, $\hat{p}_{MC}(s_c,\ s_d)$ depends on the network. When estimating $\hat{p}_{MC}(s_c,\ s_d)$ in the course of RDS analysis, researchers typically do not have access to the underlying network structure, so they estimate it via $\hat{p}_{RDS}(s_c,\ s_d)$, which is calculated based on the number of recruitments by individuals with key variable value s_c of individuals with key variable value s_d.

3. Bias in Sampling Without Replacement

We have shown that the underlying cause of bias due to sampling without replacement in RDS studies is the difference between RDS chains and those that would be created under chain-referral sampling with replacement. The next question is the magnitude and direction of that bias.

Let us begin by making an observation: sampling without replacement produces a bias in transition probability estimates when a participant in an RDS study with value $L(s_c)$ attempts to recruit another individual, say with value $L(s_d)$, but finds that individual has already been recruited. Bias occurs in this case because the equilibrium transition probabilities are based on the number of network connections between individuals with $L(s_c)$ and $L(s_d)$, but the recruitment failure prevents one of those connections from being included in the estimate counts. We make this observation formal by defining a *repeated sampling event*:

Definition 3.1 *A repeated sampling event ω is an event where participant r_i in an RDS study attempts to recruit participant r_j but discovers that r_j has already been recruited.*

Equivalently, we can think of a repeated sampling event as introducing a difference between an RDS chain and a with-replacement chain-referral sample on the sample population, with the same seeds. Now we can describe the bias due to sampling without replacement in terms of a frequency of discrete events.

Before we proceed with the rest of the analysis, it is important to point out that for the vast majority of RDS studies, we cannot measure the frequency of repeated sampling events directly, since most RDS studies do not ask recruiters how many peers they attempted to recruit into the study, nor how many of those peers refused because they had already participated. However, we can make general observations about the frequency of repeated sampling events in RDS, and, based on these observations, demonstrate analytically the dynamics of this frequency for different values of sampling (fraction, homophily, and so on).

We begin by observing that the occurrence of repeated sampling events is determined by exactly three factors:

- Network structure
- Sampling fraction
- Probability of any RDS recruitment chain following a particular edge in the network

For example, consider an undirected tree network with the RDS seed as the root. Then, regardless of sampling fraction or the probability of following any particular edge in the network, the only possible repeated sampling events are those where a recruit directly attempts to recruit her recruiter. Assuming these "backtracking" events do not occur (as we do below), no repeated sampling events occur, and the bias from sampling without replacement is zero. Conversely, consider a population where for some reason every recruit will attempt to recruit her recruiter. Then, regardless of (nonzero) sampling fraction and network structure, there will be some repeated sampling events, and the bias from sampling without replacement is nonzero.

Neither of these scenarios is likely to occur in an empirical RDS study; however, they are useful for two reasons. First, they provide us with theoretical bounds for the space of repeated sampling events. Second, they do resemble some empirical RDS scenarios. For example, networks of novice drug users in NY have been shown to resemble a star shape, with several recreational users connected to no one but a central active supplier (Wallace 1991).

In the following analysis, we will investigate first the density of repeated sampling events, and then the effect of this density on bias. We first investigate the effect of sampling fraction and network structure on the occurrence of these events, and then move on to the last factor, the probability of RDS recruitment along particular edges in the network.

Density of Repeated Sampling Events

The key factor in measuring and accounting for bias in sampling without replacement is the density of repeated sampling events, which we will call ρ_ω. We begin with a definition of ρ_ω:

Definition 3.2 *The density ρ_ω of repeated sampling events in a particular RDS study or simulation is the frequency of repeated sampling events divided by the total number of recruitment events in the study or simulation.*

A formal analysis of this quantity yields surprising observations about the correspondence between ρ_ω and particular network structures. We can use these observations to infer backwards from our understanding of network structure in hidden populations to expected levels of bias due to repeated sampling events. We begin this analysis by modifying the original assumptions about the RDS process outlined in Section 1, so as to incorporate sampling without replacement. Below, we present the new set of assumptions about RDS necessary for our analysis:

Assumption 3.3 The target population's network must be dense enough for the population to form a single component, so every node is reachable from every other node.

Assumption 3.4 Recruiters know one another, as acquaintances, friends, or those closer than friends, so their relationships are reciprocal and recruitment can occur in either direction along the tie.

Assumption 3.5 Respondents recruit as though they are selecting randomly from their neighborhoods.

Assumption 3.6 Recruits cannot attempt to recruit their recruiters. In other words, the random walk that generates the sample does not backtrack.

Assumption 3.7 The RDS recruitment process is asynchronous, that is, at no point is a potential recruit simultaneously approached by two or more recruiters.

Assumption 3.8 The RDS process begins with one seed.

Assumption 3.9 Respondents attempt to recruit a constant number k of their neighbors, or all of their neighbors, whichever is smaller. This number includes failed attempts to recruit due to repeated sampling events. *Attempt* here means that a recruiter will try to recruit some individual unless she has already been recruited, in which case the recruiter tries to recruit another individual in their network neighborhood and so on, until the recruiter has tried to recruit k individuals. The inclusion of failed attempts in k may mean that the RDS process stops when no recruiter has a legal recruit. In formal analysis, we are not concerned with the termination of specific RDS chains unless it happens deterministically, which is not the case for sampling fractions $< 100\%$. In simulations, chain termination is a concern, which we address by introducing additional seeds, one at a time (see the simulation section below).

Assumptions 3.3 and 3.4 pertain to the structure of the graph and are therefore scope conditions.

Assumptions 3.5, 3.6 and 3.7 pertain to the nature of the recruitment process, specifying a nonbacktracking random walk. Lee et al. (2012) showed that a nonbacktracking random walk retains the Markov property and will have the same stationary distribution as a simple random walk. Even though backtracking would create a repeated sampling event, Lee et al. show that eliminating backtracking does not change or bias the estimation of transition probabilities or the stationary distribution. We focus only on repeated sampling events that might produce bias estimates by excluding backtracking from our analysis.

Assumptions 3.8 and 3.9 are less reflective of empirical RDS studies, and we relax them in a simulation framework in Section 4.

In the following analysis, we focus on the graph of *potential recruits* G, in contrast to the recruitment graph of relationships between *actual* recruits. G is meant to represent the wider community, from which the seeds and the recruits are drawn. For example, in a study of jazz musicians in New York City, G is the graph of all jazz musicians in New York City. A graph consists of nodes and edges; in this example, the nodes are the jazz musicians in New York City and the edges are connections (friendships, professional relationships, etc.) between jazz musicians, along which recruitment may occur. Since the target sampling fraction, σ, is a rational number and the sample size has to be an integer (number of individuals), we define sample size as the greatest integer less than or equal to the sample fraction multiplied by the population size or $\lfloor \sigma N \rfloor$.

We now restate our observation about tree structures and the absence of repeated sampling events as a formal lemma:

Lemma 3.10 *Given assumptions 3.3–3.9, sampling without replacement cannot occur only in populations where the structure of relationships among members of the population is an undirected tree.*

Proof: Consider graph G, which is not an undirected tree graph. An undirected tree is a type of network where the edges are undirected and any two nodes are connected by exactly one path. If G is not an undirected tree, then there is at least one cycle in G that is a sequence of connected nodes (seed and recruits) $r_1 \ldots r_l$ where l is the length of the cycle and r_l has an edge to r_1. Then it is possible for r_1 to be a seed, and recruit r_2, who recruits r_3 and so on until r_l is recruited. Then it is possible for r_l to attempt to recruit r_1 at which point a repeated sampling event will occur.

Similarly, consider a graph G' that is an undirected tree graph. Consider some seed r_1. Then for any potential recruiter r consider the set of potential recruits PS. A repeated sampling event can occur only if some $p \in PS$ has already been recruited. But that means that a path exists from a seed r_1 to p that does not go through r. Since a path already exists from r_1 to p that does go through r, this means that a cycle must exist in G', which contradicts the claim that G' is an undirected tree. □

Lemma 3.10 shows that specific network structures imply particular levels of bias due to sampling without replacement. However, this does not mean there is a deterministic relationship between network structure and level of bias due to sampling without replacement, as we show with the following negative result:

Lemma 3.11 *Given assumptions 3.3–3.9, and a particular chain of recruitments, it is possible that this chain could have arisen without any repeated sampling events regardless of the underlying structure of relationships between the recruits.*

Proof: Given any connected graph G of potential recruits, we can remove edges from G until no cycles exist but all the nodes are still connected. This is the minimum spanning tree of G. Let the number of coupons for an RDS study on this network be greater than the degree of any node in the minimum spanning tree. Under these conditions, an RDS process can start at any node in G and end by recruiting all potential recruits avoiding any repeated sampling events simply by following the minimum spanning tree. □

However, Lemma 3.11 does not preclude estimation of bias due to sampling without replacement from network structure. We may not be able to calculate an exact amount of bias for a particular network structure analytically, but we can nevertheless define bounds for this type of bias and formalize its relationship to key variables such as the sampling fraction. In particular, we outline and then prove a number of theorems about the relationship between network structure and density ρ_ω of repeated sampling events. We begin with a theorem for making formal statements about the density of repeated sampling events for any network structure. This theorem will serve as a framework for proving statements about specific network structures.

Theorem 3.12 *Given assumptions 3.3–3.9, a graph G with N nodes and one or more cycles, a further assumption that each recruiter attempts to recruit exactly k neighbors, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is:*

$$\rho_\omega = \frac{\sum_{i=1}^{NR} f(r_i)}{NR} \tag{3}$$

where $f(r_i)$ is a function for recruiter r expressing the fraction of her k recruits that have already been recruited, and NR is the number of recruiters in the RDS sample.

Proof: The density of repeated sampling events is the ratio of repeated sampling events (RSE) to the total number of recruitment events (RE). Symbolically, let us represent it as:

$$\rho_\omega = \frac{RSE}{RE} \tag{4}$$

As per the statement of the theorem, we make a further assumption that recruiters make exactly k recruitment attempts—in other words, that Assumption 3.9 holds and furthermore every individual has degree at least k. This assumption greatly simplifies the analysis, and we relax it in the simulation section. With this assumption, we can rewrite ρ_ω as:

$$\rho_\omega = \frac{RSE}{kNR} \tag{5}$$

where NR is the number of recruiters in the RDS sample. For the numerator, let $f(r_i)$ be a function for recruiter r_i expressing the fraction of her k recruits that have already been recruited. Then the numerator is:

$$RSE = \sum_{i=1}^{NR} kf(r_i) \tag{6}$$

Substituting in RSE, taking the constant k out of the sum and canceling, we get:

$$\rho_\omega = \frac{\sum_{i=1}^{NR} f(r_i)}{NR} \tag{7}$$

□

Having shown a general relationship between ρ_ω and sampling fraction σ for some graph G, we proceed to show specific instances of this relationship on Poisson Random Graphs, Small-World Graphs and Preferential Attachment Graphs.

A Poisson Random Graph is a graph where all ties are randomly targeted and the nodes have a Poisson degree distribution. We use the Erdős-Rényi version of a Poisson Random Graph (Erdős and Rényi 1959).

A Small-World Graph is a graph whose nodes are embedded in a regular lattice, but a fraction of the edges between these nodes are randomly rewired, creating enough shortcuts in the graph to lead to a small graph diameter (a small world). Nodes in Small-World Graphs have a regular degree distribution, that is, all nodes have identical degree. Watts and Strogatz (1998) describe the construction and properties of Small-World Graphs.

A Preferential Attachment Graph is a graph where nodes connect to others preferentially based on their degree. Nodes in Preferential Attachment Graphs have a power-law degree distribution. Preferential Attachment Graphs are described in Barabasi and Albert (1999).

Theorem 3.13 *Given 3.3–3.9, a Poisson Random Graph G with N nodes and one or more cycles, and an RDS process with sampling fraction σ, the density ρ_ω is bounded by the following inequality:*

$$\frac{\sigma}{2(k+1)} - \frac{1}{2N} \leq \rho_\omega \leq \frac{\sigma}{2} - \frac{1}{2N} \tag{8}$$

Proof: For a Poisson Random Graph, all ties are randomly targeted, so the probability of a tie targeting an already-recruited node is given by *ν/N* where *ν* is the current number of recruits. For recruiter r_i, $\nu = i - 1$, so we can rewrite Equation 3 as follows:

$$\rho_\omega = \frac{\sum_{i=1}^{NR} \frac{i-1}{N}}{NR} \tag{9}$$

or

$$\rho_\omega = \frac{(NR - 1)}{2N} \tag{10}$$

Now we have the equation strictly in terms of *NR* the number of recruiters and *N* the population size. We can bound the number of recruiters by the following argument: In the simulation design of RDS, and also in RDS empirical studies, if an individual fails to recruit *k* recruits, a new recruiter is added. As we discuss above, we can assume for the purposes of this section that every individual has at least degree *k*, so the only way a recruiter fails to recruit *k* recruits is through a repeated sampling event, when the recruiter tries to recruit some individual who has already been recruited. So the minimum number of recruiters occurs when no repeated sampling events occur. In this case, every recruiter recruits exactly *k* recruits. We know the total number of participants in the sample (recruiters + recruits) is $\lfloor \sigma N \rfloor$, so we can derive the lower bound for the number of recruiters by solving:

$$NR + kNR \geq \lfloor \sigma N \rfloor \tag{11}$$

or

$$NR \geq \frac{\lfloor \sigma N \rfloor}{k+1} \tag{12}$$

Now let us consider the maximum number of recruiters. This occurs when all sampling events are repeated sampling events; when all current recruiter attempts lead to repeated sampling events in an empirical RDS study, a new recruiter is added to the sample. Thus, in this case, a new recruiter is added to the sample after every recruiter makes all of their recruitment attempts. This extremely rare situation would occur if all the initial seeds in an RDS study tried to recruit each other and only each other, and every subsequently added recruiter tried to recruit only from among the seeds. In this case, the number of recruiters is just the sample size, so the other side of the inequality is:

$$NR \leq \lfloor \sigma N \rfloor \tag{13}$$

Using this inequality, we can put bounds on ρ_ω as follows:

$$\frac{\frac{\lfloor \sigma N \rfloor}{k+1} - 1}{2N} \leq \rho_\omega \leq \frac{\lfloor \sigma N \rfloor - 1}{2N} \tag{14}$$

or

$$\frac{\sigma}{2(k+1)} - \frac{1}{2N} \leq \rho_\omega \leq \frac{\sigma}{2} - \frac{1}{2N} \tag{15}$$

□

Equation 14 shows that ρ_ω increases linearly in the sampling fraction (all other terms are constant for a given RDS sample). Note that for large populations, $\frac{1}{2N}$ is negligible, so the bounds on ρ_ω are:

$$\frac{\sigma}{2(k+1)} \leq \rho_\omega \leq \frac{\sigma}{2} \tag{16}$$

Theorem 3.14 *Given assumptions 3.3–3.9, a Small-World Graph G with N nodes and one or more cycles and rewiring probability p, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is bounded by the following inequality:*

$$\begin{aligned} &\frac{p(1+1-c_1-p+c_1p)}{2(k+1)}\sigma + \frac{p(1+1-c_1-p+c_1p)}{2N} - p\frac{1}{N} + c_1 - pc_1 \\ &\leq \rho_\omega \leq \\ &\frac{p(1+1-c_1-p+c_1p)}{2}\sigma + \frac{p(1+1-c_1-p+c_1p)}{2N} - p\frac{1}{N} + c_1 - pc_1 \end{aligned} \tag{17}$$

Proof: For a Small-World Graph with rewiring probability p, a fraction p of all ties are randomly targeted, while the rest are embedded within a regular lattice. Given some recruiter r_i making k recruitment attempts, kp of those attempts will be reaching random targets in the network, while $k(1-p)$ of those attempts will be reaching lattice neighbors. Accordingly, $f(r_i)$ will be an interpolation between p and $1-p$.

First, let us examine what happens in the case of lattice neighbors. Some fraction c of these will already have been recruited, in one of two ways: either they were recruited by their own lattice neighbors, or they were recruited through random ties. Let us call the fraction of neighbors recruited by their own lattice neighbors c_1, and the fraction recruited through random ties c_2.

The quantity c_1 is independent of sampling fraction. To see why, consider the example of a ring lattice. Since we are only looking at individuals recruited by lattice neighbors, the recruitment set on this network will resemble a line that grows at both ends. Each new recruit r_i appears at the end of the line and always has the same neighborhood composition with respect to recruited versus nonrecruited individuals: half are already recruited (the half of r_i 's neighbors that are closer to the seed) whereas the other half are not already recruited (the exception being when all individuals are recruited and the ends of the line connect, but that lone case will not affect our estimations).

The quantity c_2 is the probability that some lattice neighbor j of r_i has already been recruited by another node k via a randomly rewired tie. For each such j, there are approximately i potential recruiters, and each recruiter can recruit j if it has a rewired tie (probability p) and it points to j (since rewired ties are random, the probability is uniform at $1/N$).

To calculate the quantity c, let us consider the processes that generate c_1 and c_2 as probabilistic events C_1 and C_2. In the equation below, $||$ is the logical OR notation. Then:

$$\mathrm{c} = (1-p)\left(C_1||C_2\right) \tag{18}$$

$$= (1-p)(1-(1-P(C_1))(1-P(C_2))) \tag{19}$$

$$= (1-p)\left(1-(1-c_1)\left(1-p/N\right)^i\right) \tag{20}$$

$$\approx (1-p)\left(1-(1-c_1)\left(1-pi/N\right)\right) \tag{21}$$

Note that Equation 21 is an approximation, based on a derivation by Tillé (2006). This approximation holds for $pi \ll N$, meaning that as long as $pi \ll N$, the left-hand side is almost exactly equal to the right-hand side. The quantity i is bounded by the number of recruiters, NR. Thus, $pi \ll N$ so long as:

$$pNR \ll N \tag{22}$$

The quantity NR itself is bounded by the sampling fraction σ, such that $NR \leq \lfloor \sigma N \rfloor$. Accordingly, the inequality holds as long as:

$$p\lfloor \sigma N \rfloor \ll N \tag{23}$$

or,

$$p\sigma \ll 1 \tag{24}$$

In the simulation section of the article we use $p = 0.2$, $0 < \sigma < 1$ so $p\sigma$ ranges between 0 and 0.2, which is significantly smaller than 1.

Next, consider the pk attempts that reach network neighbors through rewired ties. These ties are rewired at random, so the targets of those ties will be random nodes in the network. As in Theorem 3.13, the probability that any attempt reaches a node that has already been recruited is $(i - 1)/N$ for the ith recruit. Now we are finally ready to write down ρ_ω.

$$\rho_\omega = \frac{\sum_{i=1}^{NR}\left(p\frac{i-1}{N} + (1-p)\left(1-(1-c_1)\left(1-p\frac{i}{N}\right)\right)\right)}{NR} \tag{25}$$

or

$$\rho_\omega = \frac{\frac{p(2-c_1-p+c_1p)}{N}\sum_{i=1}^{NR}(i) - p\frac{NR}{N} + c_1NR - pc_1NR}{NR} \tag{26}$$

or

$$\rho_\omega = \frac{p(2-c_1-p+c_1p)}{N}\frac{(NR+1)}{2} - p\frac{1}{N} + c_1 - pc_1 \tag{27}$$

As we showed above, the number of recruiters is between $\frac{\lfloor\sigma N\rfloor}{k+1}$ and $\lfloor\sigma N\rfloor$, so we can write:

$$\begin{aligned}&\frac{p\big(2-c_1-p+c_1p\big)}{N}\frac{\left(\frac{\lfloor\sigma N\rfloor}{k+1}+1\right)}{2} - \frac{p}{N} + c_1 - pc_1\\ &\leq \rho_\omega \leq\\ &\frac{p\big(2-c_1-p+c_1p\big)}{N}\frac{\big(\lfloor\sigma N\rfloor+1\big)}{2} - \frac{p}{N} + c_1 - pc_1\end{aligned} \tag{28}$$

or

$$\begin{aligned}&\frac{p(2-c_1-p+c_1p)}{2(k+1)}\sigma + \frac{p(2-c_1-p+c_1p)}{2N} - \frac{p}{N} + c_1 - pc_1\\ &\leq \rho_\omega \leq\\ &\frac{p(2-c_1-p+c_1p)}{2}\sigma + \frac{p(2-c_1-p+c_1p)}{2N} - \frac{p}{N} + c_1 - pc_1\end{aligned} \tag{29}$$

□

This is a much more complex form than Equation 14, but again, all the terms except for σ are constants for a particular RDS sample, so again ρ_ω increases linearly in σ.

Furthermore, recall that $p < 1$ and $c_1 < 1$. So, $p(2-c_1-p+c_1p) < 2$ and for large populations:

$$\frac{p(2-c_1-p+c_1p)}{2N} \approx 0 \tag{30}$$

or

$$\frac{p}{2N} \approx 0 \tag{31}$$

Then, for large populations, the bounds are:

$$\frac{p(2-c_1-p+c_1p)}{2(k+1)}\sigma + c_1 - pc_1 \leq \rho_\omega \leq \frac{p(2-c_1-p+c_1p)}{2}\sigma + c_1 - pc_1 \tag{32}$$

Theorem 3.15 *Given Assumptions 3.3–3.9, a Preferential Attachment Graph G with N nodes, one or more cycles, a degree distribution approximated by P(x) ≈ $x^{-\alpha}$ and rewiring probability p, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is a nonlinear function of σ that is sublinear for small values of σ and approaches linearity in σ for larger values of the sampling fraction.*

Proof: In a Preferential Attachment Graph, ties are not targeted randomly, but according to the degree of the target, with higher-degree nodes more likely to be tie targets. For recruit r_i, $f(r_i)$ thus depends not only on the number of individuals already recruited, but also on their sum degree. Specifically:

$$f(r_i) = \frac{SRF(i)}{SN} \tag{33}$$

where $SRF(i)$ is the sum degree over all nodes already recruited whereas SN is the sum degree over all nodes in the graph. In the beginning of the recruitment process, the recruited sample is more likely to contain network hubs than low-degree nodes, as all nodes (including the seed) are preferentially more likely to have ties to higher-degree nodes than to lower-degree nodes. However, the number of such hubs is very small, so the recruited sample quickly exhausts them and moves on towards lower-degree nodes. As this happens, the average degree over all recruits approaches the average degree over all nodes in the graph. When the average degree over all recruits is approximately equal to the average degree over all N nodes:

$$\frac{SRF(i)}{i} \approx \frac{SN}{N} \tag{34}$$

we have:

$$f(r_i) = \frac{SRF(i)}{SN} \tag{35}$$

$$= \frac{\frac{SRF(i)i}{i}}{\frac{N\,SN}{N}} \tag{36}$$

$$= \frac{i}{N}\frac{\frac{SRF(i)}{i}}{\frac{SN}{N}} \tag{37}$$

$$\approx \frac{i}{N} \tag{38}$$

This is almost the same expression as $f(r_i)$ for a Poisson Random Graph, where ρ_ω is linear in σ. Accordingly, in the limit of large sampling fraction, ρ_ω approaches a linear function of σ. However, for small sampling fractions, the sample may never reach this stage. In that case, the average degree over all recruits is much bigger than the average degree over all N nodes:

$$\frac{SRF(i)}{i} \gg \frac{SN}{N} \tag{39}$$

In this case, $f(r_i)$ is much bigger than $f(r_i)$ for a Poisson Random Graph. Therefore, ρ_ω values grow more quickly than for a Poisson Random Graph for small sampling fractions, but then grow ever slower as sampling fraction increases, approaching a linear growth rate. Thus, the second derivative of ρ_ω is initially negative, and it grows sublinearly for small sampling fractions. □

To give some idea of the range of sampling fractions, over which ρ_ω grows sublinearly, we consider the probability of high-degree nodes being picked in the sample, which also gives us the expected point at which these high-degree nodes are exhausted. Let us focus on nodes with above-average degree—so long as these nodes are picked, the average degree over the recruit set remains higher than the average degree over all nodes. The probability $p(n > \mu_d)$ of any one tie targeting a node n with above-average degree is given by:

$$p(n > \mu_d) = \frac{\sum_j d(j) > \mu_d}{SN} \tag{40}$$

$$\approx \frac{\int_{\mu_d}^{M} xP(x)dx}{\int_{m}^{M} xP(x)dx} \tag{41}$$

where $d(j)$ is the degree of node j, μ_d is the mean degree, M the max degree and m the min degree of G, x is degree, and $P(x)$ is the degree distribution of G. The approximation in Equation 41 is a smoothing out of Equation 40, since the degree distribution of G ranges only over discrete values of x. As we note in the theorem statement, $P(x) \approx x^{-\alpha}$. Therefore, Equation 41 evaluates to:

$$p(n > \mu_d) \approx \frac{M^{2-\alpha} - {\mu_d}^{2-\alpha}}{M^{2-\alpha} - m^{2-\alpha}} \tag{42}$$

where α is the best-fit exponent of the degree distribution of G. For $\alpha > 2$, μ_d is well-defined and equal to:

$$\mu_d = m\frac{\alpha - 1}{\alpha - 2} \tag{43}$$

So we can rewrite above as:

$$p(n > \mu_d) \approx \frac{M^{2-\alpha} - \left(m\frac{\alpha - 1}{\alpha - 2}\right)^{2-\alpha}}{M^{2-\alpha} - m^{2-\alpha}} \tag{44}$$

Note that for $\alpha > 2$, $M^{-2\alpha}$ is very close to 0. We can use that to reapproximate $p(n > \mu_d)$ as:

$$p(n > \mu_d) \approx \left(\frac{\alpha - 1}{\alpha - 2}\right)^{2-\alpha} \tag{45}$$

This function decreases superlinearly in α, but between $\alpha = 2$ and $\alpha = 3$ (the range for Preferential Attachment Graphs), it varies between .8 and .5. Now consider the fraction of nodes that have above-average degree, $P(N > \mu_d)$, which is derived from the cumulative degree distribution of *G,* which is the integral of the partial degree distribution of G, $P(x)$:

$$P(N > \mu_d) \approx \frac{\int_{\mu_d}^{M} P(x)dx}{\int_{m}^{M} P(x)dx} \approx \left(\frac{\mu_d}{m}\right)^{-\alpha+1} \tag{46}$$

$$= \left(\frac{\alpha-1}{\alpha-2}\right)^{1-\alpha} \tag{47}$$

$$= p(x > \mu_d) \cdot \frac{\alpha-2}{\alpha-1} \tag{48}$$

In other words, given exponent α of 2.1, about 80% of the ties will be targeting nodes with above-average degree, whereas only about 10% of the nodes will have above-average degree. This disparity suggests that a sublinear relationship will exist between ρ_ω and σ, given α of 2.1 and sampling fractions much lower than ten percent. We can establish similar relationships for other values of α and σ, but note that as α increases, fewer and fewer of the early ties will point to above-average degree nodes. A complex nonmonotonic relationship therefore exists between the power-law exponent and the relationship between sampling fraction and density of repeated sampling events.

Bias Due to Repeated Sampling Events

We formalize the relationship between the density of repeated sampling events ρ_ω and the bias due to sampling without replacement. This bias can be expressed as the difference between the equilibrium transition probabilities for a Markov Process modeling a chain-referral sample with replacement and the estimated transition probabilities between different groups in an empirical RDS sample:

$$Bias_{SWOR} = \sum_{s_i}\sum_{s_j}\left|p_{MC}(s_i, s_j) - \hat{p}_{RDS}(s_i, s_j)\right| \tag{49}$$

where s_i, s_j are different values of the key variable L analyzed in the course of an RDS study as described in Section 2, and $\hat{p}_{RDS}(s_i, s_j)$ and $p_{MC}(s_i, s_j)$ are defined in Equations 1 and 2, respectively. In the case where no repeated sampling events are possible (e.g., on an undirected tree as described in Lemma 3.10), this bias tends asymptotically to 0 in the sampling fraction σ:

$$\lim_{\sigma \to 1}(Bias_{SWOR}) = 0 \tag{50}$$

In the case where repeated sampling events are possible, each event initially introduces a small amount of bias. Recall that $Rec(A,B)$ is the collection of all recruitment events where an individual in A recruits another individual in B, $D(A,B)$ is the set of network connections between individuals in collection A and individuals in collection B and the definitions of $\hat{p}_{RDS}$ and $\hat{p}_{MC}$ given in Equations 1 and 2. Consider a single repeated sampling event in the sampling process where a recruiter $r_{1,}$ $L(r_1) = s_1$ tries to recruit another individual $r_{2,}$ $L(r_2) = s_2$, but r_2 has already been recruited.

In the limit of $\sigma \longrightarrow 1$, the number of recruitments from s_1 nodes to s_2 approaches the number of edges between s_1 and s_2 minus the single failed recruitment.

$$\begin{aligned} &\left|Rec\left(\langle r_i :\ L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right| \\ &\longrightarrow \left|D\left(\langle r_i : L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right| - 1 \end{aligned} \tag{51}$$

At the same time, the number of recruitments from s_1 nodes to any other node approaches the number of edges between s_1 and all other nodes minus the single failed recruitment:

$$\left|Rec\left(\langle r_i :\ L(r_i) = s_1\rangle, R\right)\right| \longrightarrow \left|D\left(\langle r_i : L(r_i) = s_1\rangle, R\right)\right| - 1 \tag{52}$$

Let $a = \left|D\left(\langle r_i : L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right|$ and $b = \left|D\left(\langle r_i :\ L(r_i) = s_1\rangle, R\right)\right|$ so

$$\hat{p}_{RDS}(s_i, s_j) \longrightarrow \frac{a-1}{b-1} \tag{53}$$

With a and b as defined above, $\hat{p}_{MC}(s_i, s_j) = \frac{a}{b}$ and so $Bias_{SWOR} \longrightarrow \frac{a}{b} - \frac{a-1}{b-1}$. In general, if the density of repeated sampling events ρ_ω is uniform across all nodes and each node makes the same number of recruitment attempts, then, in the limit of $\sigma \longrightarrow 1$:

$$\hat{p}_{RDS}(s_i, s_j) \longrightarrow p_{MC}(s_i, s_j)(1 - \rho_\omega) \tag{54}$$

for all s_i, s_j and so $Bias_{SWOR} \longrightarrow \sum_{s_i} \sum_{s_j} \left|p_{MC}(s_i, s_j)(\rho_\omega)\right|$. Note that the assumption that each node makes the same number of recruitment attempts is not entirely unrealistic in empirical RDS studies, since the number of coupons per recruiter is usually capped at a small value. The other assumption, that repeated sampling event density is uniform across all nodes, is less realistic, but useful for generalized results across network structures. In the simulation section of this article, we relax the uniform density assumption.

Bias Due to Degree Differential and Group Size

The case explored in Theorem 3.15 suggests that the degree of recruits can play an important role in creating bias due to sampling without replacement. In this section, we consider a scenario wherein two groups of recruiters are present, one with a drastically higher degree than the other, and explore the implications for density ρ_ω of repeated sampling events. We also consider the effect of group size on bias, that is, the case where two groups are present in the target population, but one group has many more members than the other. In this section, we make a further simplifying assumption for RDS samples: when a target population is divided into two groups, all members of a group have the same degree. This is a strong assumption, but it helps illustrate the fundamental effect of degree differential and group size on bias. We relax this assumption in the simulation section below.

The networks we construct in this section have the property that the probability of an edge targeting a particular node $= td(i)$ where $t > 0$ is a constant and $d(i)$ is the target node's degree. This mode of network construction allows us to examine a simplified version of preferential attachment tie formation behavior found in many empirical networks. Given our earlier assumption that all nodes in one group have the same degree, we cannot construct actual preferential attachment networks; again, we relax this constraint in the simulation section below, where we examine true preferential attachment networks.

First, we explore only the effect of degree differential. Consider a population of potential recruits that consists of two equal-sized groups *A* and *B*, embedded in a network as described above. Furthermore, we have:

$$\mu_d(A) = c\mu_d(B) \tag{55}$$

where $\mu_d(A)$ is the mean degree of group *A* and $\mu_d(B)$ is the mean degree of group *B*, and *c* is a constant. Given a uniform distribution of recruits between network neighbors as per Assumption 3.5, the probability of a recruit coming from group *A* is *c* times the probability of a recruit coming from group *B*. Then, given *r* recruits, under sampling with replacement, $rc/(c + 1)$ of them will come from group *A* and $r/(c + 1)$ of them will come from group *B*. However, the sample is collected without replacement, so the sampled proportions of recruits from group *A* and *B* will differ from the with-replacement condition.

We illustrate the difference in sampled proportions of recruits from *A* and *B* under conditions of sampling without replacement with the following toy scenario: consider a population of 100 individuals, 50 of which are in group *A* and the other 50 in *B*. The sampling fraction is 70 percent and the average degree of group *A* is six times the average degree of group *B*.

Under conditions of sampling with replacement, 70 individuals are recruited (some multiple times), 60 of those individuals are from group *A* and ten from group *B*. However, under conditions of sampling without replacement, all we know is there are 60 recruitment attempts targeted at group *A* and ten at group *B*. Some of these attempts may lead to repeated sampling events, where an individual in group *A* has already been recruited, and others do not. An estimate of the density of repeated sampling events ρ_ω in these scenarios depends on calculating the expected fraction of recruitment attempts that end up as repeated sampling events.

We begin with a simple observation: consider the situation that, under sampling without replacement, the first 50 recruitment attempts targeted at group *A* each target a distinct recruit. Then we know the last ten attempts targeted at group *A* automatically lead to a repeated sampling event. From this situation, we observe that at some point a group may become *exhausted,* after which point all recruitment attempts targeting that group automatically lead to repeated sampling events. After this point, the density ρ_ω of repeated sampling events becomes an interpolation between 1 (the rate for events that target group *A*) and whatever the rate was previously; this interpolation rapidly converges to 1 as sampling fraction increases. We present a formal proof of this observation below.

We now investigate the case when *A* is not exhausted. In this case, recruitment attempts targeted at *A* lead to a repeated sampling event with a probability that rises in the number of individuals already recruited from *A*. That probability is zero if no recruits have yet come from *A* and approaches unity as *A* approaches exhaustion.

Finally, consider the effect of group size, which is very simple: the smaller the size of *A,* the more quickly it approaches exhaustion. In other words, the smaller *A* is, the earlier the onset of the exhausted regime, during which ρ_ω converges rapidly to one as sampling fraction increases.

We now combine these observations into a formal argument. We begin by proving a lemma that gives a formal expression for the expected number of distinct recruits from some group U given m recruitment attempts targeting U. This lemma is necessary to calculate the exhaustion point for groups, and relies on the assumption we make at the beginning of this section: all individuals in a particular group have the same degree. We then use the lemma to prove a "master equation" theorem that combines sampling fraction, degree differential and group size into one expression for ρ_ω.

Lemma 3.16 *Given a group U of size N that has not yet been exhausted, such that all individuals in the group have the same degree, the expected number of distinct recruits from U given m recruitment attempts targeting U is bounded by the inequality:*

$$m - \frac{m(m-1)}{2N} < DR < m \tag{56}$$

Proof: The quantity DR is equivalent to the number of distinct elements NDE after sampling m elements with replacement from a set of N elements with uniform selection probabilities, which is given in Tillé (2006):

$$NDE = N - \frac{(N-1)^m N!}{N^m (N-1)!} \tag{57}$$

or

$$NDE = N\left(1 - \left[\frac{N-1}{N}\right]^m\right) \tag{58}$$

Focusing on the exponential term in Expression 58, we have:

$$\left[\frac{N-1}{N}\right]^m = \left[1 - \frac{1}{N}\right]^m = \left[1 + \frac{-1}{N}\right]^m = \tag{59}$$

by binomial expansion:

$$\begin{aligned} &= \sum_{k=0}^{m} \binom{m}{k} \left[\frac{-1}{N}\right]^k = \binom{m}{0} 1 + \binom{m}{1} \frac{-1}{N} + \binom{m}{2} \frac{1}{N^2} \\ &\quad + \ldots + \binom{m}{m} \left[\frac{-1}{N}\right]^m \end{aligned} \tag{60}$$

This series has the property that, for any $k \leq m$, $m < N$ the $k+1$st element is smaller in magnitude and opposite in sign to the kth element. The sign opposition comes from the -1 in the power term of the series. The magnitude difference comes from the fact that the $k+1$st element is $O([m/N]^k)$, which decreases in k since $m < N$.

This property implies that the first few terms will dominate the series. In particular, we can establish bounds of the series with the second and third partial sums: $1 - m/N$ and $1 - \frac{m}{N} + \frac{m(m-1)}{2N^2}$. Every subsequent term will alternatively drive the series closer to

$1 - m/N$ and to $1 - \frac{m}{N} + \frac{m(m-1)}{2N^2}$, by an ever-decreasing degree, so the final sum will always stay within those bounds. Accordingly, we can approximate the inner term as follows:

$$1 - \frac{m}{N} + \frac{m(m-1)}{2N^2} > \left[\frac{N-1}{N}\right]^m > 1 - \frac{m}{N} \tag{61}$$

We can now rewrite Equation 56 as:

$$N\left(1 - 1 + \frac{m}{N} - \frac{m(m-1)}{2N^2}\right) < NDE < N\left(1 - \left[1 - \frac{m}{N}\right]\right) \tag{62}$$

or

$$m - \frac{m(m-1)}{2N} < NDE < m \tag{63}$$

□

What does Equation 63 tell us? Instead of targeting m distinct nodes, m recruitment attempts target some slightly smaller number $m - \epsilon$ nodes. In other words, $m - \epsilon$ ties target distinct nodes in the network, and the remaining ϵ ties are redundant, that is, lead to repeated sampling events.

Operationally, we can approximate DR by setting NDE to its lower bound (by the argument above, NDE will be much closer to its lower bound than to its upper bound):

$$\epsilon = \frac{m(m-1)}{2N^2} \tag{64}$$

Then:

$$DR = NDE \approx m - \frac{m(m-1)}{2N^2} \tag{65}$$

We now follow with a definition of group exhaustion and then the "master equation" theorem. In Theorem 3.18, we use big O notation written as ($f = O(g)$), where f and g are non-negative functions. This notation indicates that f is asymptotically upper bounded by g, in other words, that there exists an integer n_0 and a constant $c > 0$, such that for all integers $n > n_0$, $f(n) \leq cg(n)$. In this particular case, we claim that ρ_ω is asymptotically bounded by another expression, either σ if neither group is exhausted or an expression that tends asymptotically to 1 as $\sigma \longrightarrow 1$ if one of the groups is exhausted. In the latter case, Theorem 3.18 shows that ρ_ω is always less than some function, and that function itself is always less than 1.0 but approaches it quickly as σ increases.

Definition 3.17 *A group of potential recruits is said to be exhausted if, in the course of an RDS recruitment process, every individual in that group is recruited.*

Theorem 3.18 *Given Assumptions 3.3–3.9, and a population of N potential recruits split into two groups A and B, with α individuals in A and β individuals in B, such that every individual in A has degree d(A) and every individual in B has degree d(B), and the probability of an edge targeting a particular node $= kd(i)$ where $k > 0$ is a constant and d(i) the node's degree, at sampling fraction σ the frequency of repeated sampling events,*

ρ_ω, is approximated by:

$$\rho_\omega = O(\sigma) \textit{ if neither group is exhausted} \tag{66}$$

$$\rho_\omega = O\left(1 - \frac{(1-k)PER(g)}{TR(g)}\right) \textit{ if some group } g \textit{ is exhausted}, \tag{67}$$

$$\textit{which tends asymptotically towards } 1 \textit{ as } \sigma \longrightarrow 1$$

where $0 < k < 1$ is some constant, PER(g) is the number of recruitment attempts made targeting group g before g is exhausted, and TR(g) is the total number of recruitment attempts made targeting group g. Further, we can express TR(g) as:

$$TR(\mathrm{g}) = \frac{d(g)}{\Sigma_{g' \epsilon \{A,B\}} d(g')} \sigma N \tag{68}$$

And PER(g) as:

$$PER(\mathrm{g}) = s(g) + \epsilon \tag{69}$$

where $s(A) = \alpha$, $s(B) = \beta$ and ϵ is some small positive constant.
Proof: The proof follows from the observations before Lemma 3.16. For each of the two groups A and B in the target population, one of two cases is possible: either the group is exhausted, or it is not. If neither group is exhausted, then we only have to consider repeated sampling events due to the same recruit being targeted multiple times by chance. Since we assume every recruit within a group has the same degree, the probability of a repeated sampling event for some group is determined entirely by the number of individuals already recruited from that group. We can then use the same reasoning as in Theorem 3.13 to show that, when no group is exhausted, ρ_ω is dominated by the sum of two linear functions of s, which is itself a linear function of σ.

Now consider the case when some group g is exhausted. In this case, ρ_ω is an interpolation between the density of repeated sampling events prior to exhaustion, and 1, which is the density after exhaustion, and $\rho_{\omega NE}$, which is the density from any nonexhausted groups. Given TR total recruitment attempts made targeting g and PER of those attempts made prior to exhaustion, we can express this interpolation as:

$$\rho_\omega = O\left(\rho_{\omega NE} + \frac{kPER(g)}{TR(g)} + \frac{TR(g) - PER(g)}{TR(g)}\right) \tag{70}$$

or

$$\rho_\omega = O\left(\rho_{\omega NE} + 1 - \frac{(1-k)PER(g)}{TR(g)}\right) \tag{71}$$

where k is some constant between 0 and 1 representing the linear dependence between number of recruitment attempts and density of repeated sampling events prior to exhaustion. This asymptotic bound equation consists of two parts: $\rho_{\omega NE}$, which grows linearly in sampling fraction, and $1 - (1-k)PER(g)/TR(g)$. The growth of the second part depends entirely on $TR(g)$, since $PER(g)$ remains constant once g is exhausted and k is a constant. We now derive an expression for $TR(g)$. Since the network is a preferential attachment network, the number of recruitment attempts targeting group g is simply the total number of recruitment attempts times the ratio of the group degree to the sum degree:

$$TR(\mathrm{g}) = \frac{d(g)}{\Sigma_{g' \epsilon \{A,B\}} d(g')} \sigma N \tag{72}$$

Plugging Equation 72 into Equation 71, we have:

$$\rho_{\omega} = O\left(\rho_{\omega NE} + 1 - \frac{(1-k)PER(g)(d(A)+d(B))}{\sigma d(g) N}\right) \tag{73}$$

Equation 73 is dominated by the nonlinear term $1 - (1-k)PER(g)(d(A) + d(B))/(\sigma d(g)N)$ so we can rewrite it as

$$\rho_{\omega} = O\left(1 - \frac{(1-k)PER(g)(d(A)+d(B))}{\sigma d(g) N}\right) \tag{74}$$

Note that this quantity increases asymptotically towards 1 as $\sigma \longrightarrow 1$ since all the terms except σ are constants and σ is in the denominator of the fraction.

The only remaining piece of the proof is to derive an expression for $PER(g)$. $PER(g)$ is the number of recruitment attempts made targeting group g before the group is exhausted. Note that this is not simply the number of individuals in g, as we showed Lemma 3.16, making m recruitment attempts generally yields fewer than m distinct recruits. In order to calculate this value more precisely, we need to solve Approximation 65 for m given $DR = s(g)$, that is:

$$s(g) \approx m - \frac{m(m-1)}{2s(g)^2} \tag{75}$$

This equation is highly nonlinear but it is easy to estimate an approximate solution as $m \approx s(g) + \epsilon$ for some small ϵ (too small to significantly affect ρ_{ω}). Plugging in that estimate, we have:

$$s(g) \approx s(g) + \epsilon - \frac{1}{2} - \frac{(2\epsilon - 1)}{2s(g)} - \frac{\epsilon(\epsilon - 1)}{2s(g)^2} \tag{76}$$

For $s(g) \gg \epsilon$, the right-hand side is bigger than the left-hand side, and we have a sufficient condition—enough recruitment attempts have been targeted at g to exhaust it. So, we can approximate *PER(g)* as:

$$m \approx s(g) + \epsilon \tag{77}$$

□

4. Simulations

We employed simulation to explore scenarios not addressed by the analytic results. The analytic results do not specifically account for multiple seeds or recruitment processes with one versus multiple coupons given to recruiters, nor do they account for homophily, so simulating the chain-recruitment process provides additional insight into how the presence of multiple seeds and the number of coupons available to the recruiters might impact the nonreplacement bias. In these simulations, the targeted subpopulation comprised 20% of the nodes. We simulated respondent-driven sampling and calculated the Heckathorn 2007 RDS estimates for the target subgroup.

Networks

The simulated chain-recruitment samples were generated from Watts–Strogatz Small-World Networks, Barabási-Albert preferential attachment networks and Erdős-Rényi random graphs as implemented in the Python package NetworkX (Hagberg et al. 2008). These networks have uniform, power-law, and Poisson degree distributions, respectively. We selected network parameters to maintain a mean degree of about eight for standard graphs and 16 for the higher-density graphs. Networks were of orders 500 and 5000 nodes. We present results for the 5000-node networks because we were unable to obtain meaningful results for low sampling fraction on the smaller networks. For example, a five-percent sample of a 500-node network would yield only 25 individuals. At this very small scale, RDS samples are extremely idiosyncratic and dependent on the seeds, so the variance between individual RDS samples is too large to produce a consistent pattern of results. Holding the number of nodes constant while manipulating sample fraction through sample size most closely models the tradeoffs field researchers must consider when implementing RDS. We recognize that allowing sample size to covary with sampling fraction can make the results harder to interpret, so we conducted a second set of simulations, following Gile (2011), holding sample size constant while varying the number of nodes in the network to manipulate the sampling fraction. This second approach covaries network density with sampling fraction. We report results that are consistent between the two approaches and note a few differences in the results section. Furthermore, we investigated large networks of up to 20,000 nodes and found that the results for these large networks did not differ substantially from the results reported below.

To manipulate homophily in the population, we created two separate networks. One network contained the entire target subgroup and the other network contained the rest of the population. These networks were then merged using a random double-edge swap that preserves the degree of each node (Maslov and Sneppen 2002). Different levels of rewiring result in different numbers of crosscutting ties, and thus create different levels of homophily. The target subgroup population proportion was held constant at $p = .20$. The networks ranged between zero and .75 homophily, where zero corresponds to the homophily score when ties in the network are completely at random and the maximum score of 1 corresponds to a group with no ties to other groups. Details of the calculation and rationale are described in Heckathorn (2002).

To explore the effect of high and low relative mean degree for the target population, we reallocated edges from one subgroup to the other while maintaining overall mean degree and similar overall edge count. Due to the interdependence of network parameters and the particular network generation algorithms used, it was impossible to reallocate edges in these networks without also causing some variation on other parameters. We chose to hold the node count constant while matching mean degree within groups across each network type as closely as possible. We permitted slight variation in the number of total edges as long as this variation was less than one extra edge per node. For the low mean degree networks, the mean degree for the nontarget group was 1.5 times higher than the mean degree for the target group. In the high mean degree networks, the mean target group degree was 1.5 times the mean nontarget group degree. We collected simulated samples from each of the three levels of mean degree differences for each of the three network types. Table 2 shows the variation in the total number of edges for different levels of mean degree difference.

As documented in Gile (2011) and Gile and Handcock (2010), bias in RDS estimate increases with larger differences in mean degrees among groups. In other words, estimates of variables that are strongly related to degree will exhibit much larger bias than estimates of variables that have no relationship to degree. In order to select an appropriate degree ratio for our simulations, we estimated degree ratios for five public health RDS studies: two RDS surveys of Latino MSM (Ramirez-Valles et al. 2005) and three U.S. sites of the SATH-CAP studies (Iguchi et al. 2009). We examined 14 variables for each of the Latino MSM studies, resulting in 37 degree ratios per study; we examined 19 variables for each of the SATH-CAP studies, resulting in 48 degree ratios per study. We found that only three of the 218 degree ratios examined were greater than 2. Furthermore, we found that only SATH-CAP's RTI site has more than 25% of its ratios greater than 1.5. Over all studies and variables, 50% of ratios were less than 1.2, 88% of the ratios were less than 1.5, and 97.7% were less than 2. We conclude from this analysis that a significant majority of network size ratios will be less than 1.5 in public health RDS studies, and that virtually all network size ratios will be less than 2. Therefore, our simulations employ a ratio of 1.5 to be representative of the majority of ratios a typical RDS public health study will observe.

Simulation Parameters

We simulated with- and without-replacement RDS using both branching and nonbranching referral processes. Each simulation started from six randomly selected seed nodes and each seed was granted a recruitment quota of c. These seeds form the first wave of recruiters. Each recruiter recruited up to c of their available neighbors and each of these new recruits was allocated c successful recruitments for the next recruiting wave. In the without-replacement samples, nodes were considered available for recruiting only if they were not already in the sample. Each recruiter selected their recruits at random from their available neighbors until either c new recruits were generated or all available neighbors were recruited. Recruitment was asynchronous and the order of execution of recruitment privileges was randomized by node.

Table 2. Differences in Group Mean Degree for Target and Nontarget Groups

Target Group Mean	Target	Non-Target	Total Edges
Lower	12	18	42000
Same	16	16	40000
Higher	24	16	44000

The recruitment quota was one for nonbranching chains ($c = 1$) or three for branching chains ($c = 3$). We considered other quotas, including $c = 2$, $c = 4$ and c chosen uniformly at random between 0 and 4, but chose to focus on the distinction between branching and nonbranching samples as all branching samples were substantially similar and recruitment quotas are not the main focus of this article.

Each sample began with six seeds selected uniformly at random from the nodes in the network and chain recruitment continued until the target sample size was reached. In a typical RDS field study, where sampling is without replacement, the sample size is the number of people interviewed and equals the number of interviews and the number of unique nodes visited. When simulating with-replacement RDS, the sample size corresponds to the number of node visits (interviews) rather than the number of unique nodes visited. In simulated with-replacement samples the number of nonduplicate nodes in the sample ranged from about 60 to 99% of the sample size. As expected, the percentage of unique nodes visited decreased as sampling size increased. The relationship is linear with an average of 90% unique visits for sampling fractions of 0.05 and an average of 65% for sampling fractions of 0.8. The pattern was substantially similar on all three network types.

We explored sampling fractions of 0.05, 0.1, 0.2, 0.4, 0.6 and 0.8. Sample chains constructed without replacement occasionally terminate early when all nodes near the recruiting nodes have already been recruited, and the probability of early chain termination increases with the target sample fraction. To reach the target sample sizes, we adopted the following procedure: if a simulated RDS sample has no "productive" recruitment chains, but has not yet reached target sample size, add a single new seed to the sample, chosen uniformly at random from the set of nodes not yet recruited. We applied this procedure iteratively to all simulated samples that had failed to reach their target sample size, until the total number of recruits + seeds in these samples was equal to the target sample size.

Sample Filtering

The calculation of Heckathorn (2007) RDS estimates requires cross recruitment among the different subgroups in the population. When networks contain very few intergroup ties (i.e., have extreme levels of homophily), RDS samples drawn from these networks often fail to capture any of these critical ties. We excluded samples with fewer than four intergroup recruitments in each direction.

Though most samples could be collected from the initial seed cluster, some samples required the addition of new seeds as described above. While many of these samples succeeded after the addition of a few new seeds, some samples required the addition of many seeds to reach the target sampling fraction. Since seeds are recruited with an unknown probability in real-world RDS studies, their sample inclusion probability cannot be known. Instead, the network size for seeds is treated as missing data and must be imputed. In effect, this means that samples with a large proportion of seed nodes have a large proportion of missing data. When the number of seeds becomes too large, RDS estimation is inappropriate, so samples with more than five percent seeds were excluded from the analysis.

The proportion of samples excluded by these criteria depends on sampling method, sample size, and network structure. For instance, small samples from highly homophilous networks are less likely to capture cross-group ties. In less dense networks with nonbranching samples, samples were much more likely to dead end, thus requiring additional seeds. For the non-constant sample size simulations two to twelve percent of samples at sampling fraction level .05 were excluded and most of the excluded samples were drawn from the most homophilous networks. The exclusion rate rapidly decreases with sampling fraction, dropping below one percent by 20% sampling fraction. We attribute this to the small sample sizes at low sampling fractions. The constant sample size simulations ($n = 500$) had a more constant rejection rate, typically below five percent.

The two criteria for sample exclusion—fewer than four intergroup recruitments in each direction and more than five percent seeds—are based on easy-to-recognize sample characteristics that are observable by researchers conducting RDS studies in the field and reflect the importance of cross-group ties to the RDS estimation process.

Simulation Results

We now present results of simulations that confirm and extend our analytical results. The parameter space we are examining is extremely high dimensional, including sampling fraction, network structure, relative mean degree of the target group, and homophily. In this section, we first present the results about ρ_ω, which confirm our analytical results, and then describe the bias of simulated RDS samples, which extend our analytical results.

We begin by looking at ρ_ω as a function of sampling fraction for without-replacement RDS samples across different network structures and mean degree values for the target group. Figures 1, 2, and 3 correspond to Poisson Random Graphs, Small-World Graphs, and Preferential Attachment Graphs, respectively. In each figure, degree differential changes from left to right, with a target group having a smaller mean degree than the rest of the network on the left pane; same mean degree as the rest of the network in the center pane; and a greater mean degree than the rest of the network in the right pane. See Table 2 for the absolute degree values of the target and nontarget groups in both cases. Finally, within each pane, we have sampling fraction on the x-axis and frequency of repeated sampling events on the y-axis. The multiple lines for each pane correspond to the nonbranching ($c = 1$) and branching ($c = 3$) cases and the width of the 95% confidence intervals for these cases, respectively.

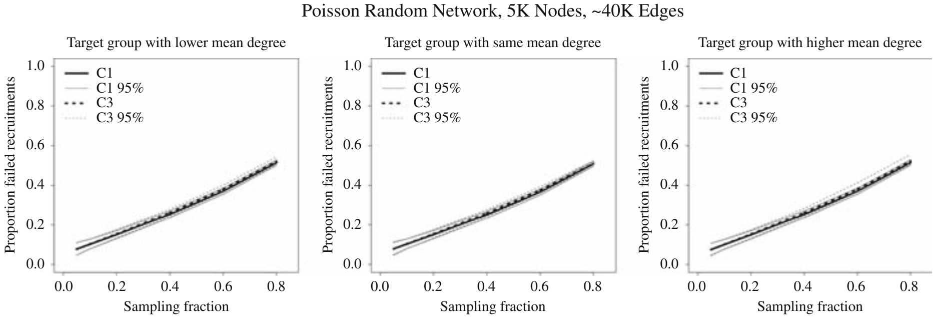


Fig. 1. Recruitment failures for Poisson Random Graphs

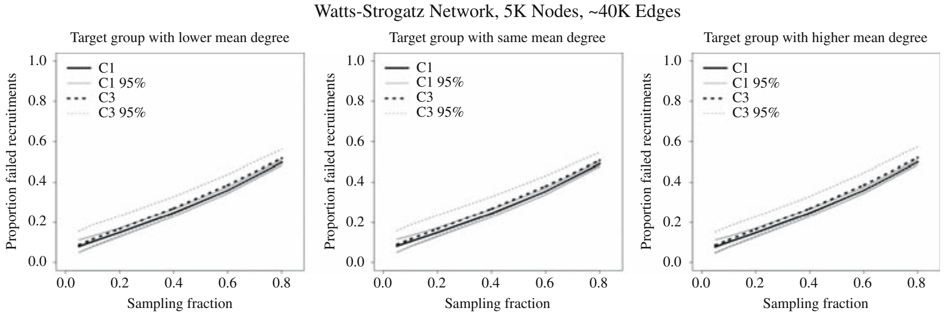


Fig. 2. Recruitment failures for Small-World Graphs

In Figure 1, we can see that in all three panes the relationship between ρ_ω and σ is linear, as predicted by Theorem 3.13. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond to Equation 66 in Theorem 3.18.

In Figure 2, we can see that in all three panes, the relationship between ρ_ω and σ is linear, as predicted by Theorem 3.14. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond with Equation 66 in Theorem 3.18.

In Figure 3, we can see that in all three panes, the relationship between ρ_ω and σ is slightly sublinear when $\sigma < .2$, and then quickly approaches linearity, as predicted by Theorem 3.15. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond with Equation 66 in Theorem 3.18.

Note that in all Figures 1–3 there is no significant difference between the branching and the nonbranching cases, which confirms the independence of our analytic results in Theorems 3.13–3.15 and 3.18 on the number of respondents recruited by each recruiter, represented by the parameter k in Theorem 3.12.

We present a set of results that show the effect of homophily on ρ_ω. Our analysis does not account for the effect of homophily on the frequency of repeated sampling events, so the simulations serve as a useful counterpart for analyzing recruitment failures in high-homophily regimes. In our work, we use the definition of homophily presented in Heckathorn (2002).

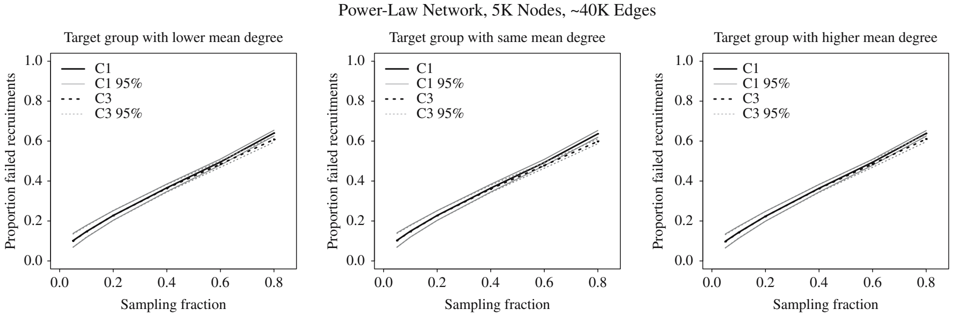


Fig. 3. Recruitment failures for Preferential Attachment Graphs

We find no significant difference in mean or the variance of the probability of repeated sampling events, ρ_ω, across the full range of homophily values we explore (0 to .75) for Poisson Random Graphs or Preferential Attachment Graphs, even when varying the target group mean degree and the branching value. For Poisson Random Graphs, ρ_ω ranges between 0.22 and 0.25 across the full range of homophily, and for Preferential Attachment Graphs, ρ_ω ranges between 0.35 and 0.36 across the full range of homophily.

Figure 4 shows the relationship between ρ_ω and homophily, averaged across all values of sampling fraction, for the Small-World Graph. In this figure, as in Figures 1–3, the degree ratio changes from left to right, with a target group having a smaller mean degree than the rest of the network on the left pane; the same mean degree as the rest of the network in the center pane; and a greater mean degree than the rest of the network in the right pane.

Figure 4 shows that for Small-World Graphs, homophily has an apparent effect on ρ_ω. Initially, the branching and nonbranching cases start out with the same level of ρ_ω, but as homophily increases, ρ_ω increases superlinearly for the branching case. The increase is likely due to the high level of clustering in Small-World Networks. More clustered networks have more within-group collisions even with a homophily of 0, and an increase in homophily will only exacerbate the within-group collisions for these networks. In contrast, Poisson Random Graphs and Preferential Attachment Graphs feature low levels of clustering, and the effect of homophily on ρ_ω is negligible ($< 5\%$) across the range of homophily values up to 0.7. Overall, the effect of homophily on nonreplacement bias is negligible compared to the effect of sampling fraction. High homophily does impact the probability of capturing cross-group ties in an RDS sample and is an important consideration in RDS survey design, but does not appear to contribute to the bias from sampling without replacement.

We now present simulation results that show the effect of sampling fraction on overall RDS bias. Sampling without replacement is only one factor that could affect the bias of an RDS estimate. For instance, in a larger sample, we would find more repeated sampling events, which may lead to increased bias, but also less dependence on initial conditions (seeds), which may lead to decreased bias. Therefore, we investigate both overall RDS bias and the part of it that is attributable to sampling without replacement. We run two parallel sets of simulations: one sampling without replacement, as above, and a second on the same network but sampling with replacement, so individuals can participate in a study more than once. For clarity of visualization, we here focus exclusively on the branching ($c = 3$) RDS samples. We discuss nonbranching samples at the end of this section.

The figures below show both mean bias and sampling variability, which we define as the range of the 95% confidence interval of the sampling distribution over the set of samples. Figure 5 shows the mean bias and sampling variability for the Poisson Random Graph. The three panels correspond to three levels of differential degree, with the target group having smaller degree on the left pane, equal degree in the center pane, and higher degree on the right pane. We use a legend to differentiate between samples drawn with (WR) and without replacement (WOR).

There are three important observations to make based on these graphs: first, while the expected bias in samples drawn with replacement is zero, the expected bias in samples drawn without replacement remains small (less than five percentage points away from the

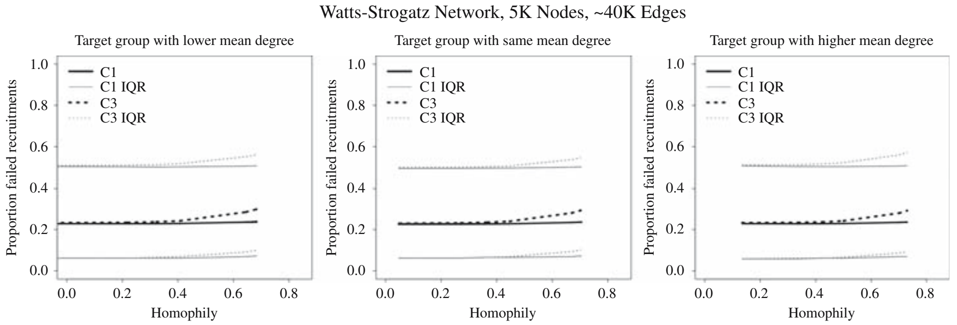


Fig. 4. Recruitment failures by homophily for Watts-Strogatz Network

true value), even up to sampling fractions of 80%. Therefore, we can say that *the effect of violating the sampling-with-replacement assumption on the expected bias of RDS estimates is five or fewer percentage points across the parameter range explored in this study*. Second, the mean bias and sampling variability for samples drawn without replacement are nearly identical to mean bias and sampling variability for samples drawn with replacement up to a sampling fraction of about 20%. Therefore, we can say that *sampling without replacement is not a major source of estimator bias in RDS studies for sampling fractions under 20%.* Finally, also across all three graph structures, we see a nonlinear relationship between the variability of samples drawn with versus without replacement. The variability of samples drawn with replacement shrinks to zero in the limit of 100% sampling fraction (not shown in graphs). The variability of samples drawn without replacement decreases and then increases: it follows the behavior of samples drawn with replacement up to sampling fraction ~40%, and then increases rapidly, except when the target group has higher mean degree. The reasons for the pattern will be explored in a future paper.

We plot the mean bias and sampling variability for two other types of networks, Watts-Strogatz and Power Law, in the figure in the supplemental data. The results for these types of networks are substantially similar to the results for Poisson Random Graphs as shown in Figure 5.

Nonbranching samples represent an idealized process that is not representative of empirical RDS studies, but the simulations indicate mean bias for nonbranching

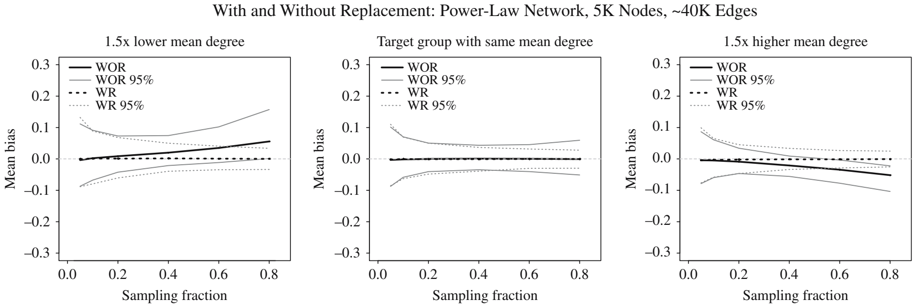


Fig. 5. Mean bias by sampling fraction on a Poisson Random Network structure for samples drawn with (WR) and without replacement (WOR)

without-replacement samples is very similar to the branching without-replacement samples (not shown in graphs). The relationship between sampling fraction and the sampling variability of nonbranching without-replacement samples is similar to that of nonbranching with-replacement samples (not shown in graphs).

Branching samples are less prone to chain exhaustion than nonbranching samples on the lower density graphs with mean degree of eight. Branching samples were able to reach their target sample size from the initial six seeds but nonbranching chains often became stuck, requiring many additional seeds to reach the target sample size on all network types. To reach the highest levels of sampling fraction, hundreds of additional seeds were required. Our sample filtering capped the allowable additional seeds at five percent of the overall sample. On the more dense mean degree sixteen networks, nonbranching samples are able grow as effectively as branching samples. Though not the main focus of this article, the observed difference in robustness between branching and nonbranching samples highlights the practical necessity of recruitment quotas greater than one for some network structures.

5. Discussion

The results above show the effect of sampling fraction, degree distribution, degree differential and group size on bias in RDS studies, both specifically due to sampling without replacement (in Figures 1−4), and overall bias from the true prevalence figures (in Figure 5). We can make six general observations based on the results:

The density of repeated sampling events, ρ_ω, increases monotonically in sampling fraction across network structures and degree differential. This increase is predicted by our analysis, and is generally linear for Poisson Random Graphs and Small-World Graphs, and slightly superlinear for Preferential Attachment Graphs. This increase suggests that bias due to sampling without replacement increases steadily with sampling fraction.

Homophily has a small effect on ρ_ω outside of Small-World Networks, which have a high level of clustering. Even for Small-World Networks, ρ_ω increases by less than ten percent from homophily $= 0$ to homophily $= .7$. Theoretical analyses (Heckathorn 2002, 28) have shown that the standard error of an RDS estimate increases exponentially with increases in homophily, so RDS is not a suitable sampling method for networks with homophily above .7.

Overall bias remains small across the range of simulated parameters, less than five percentage points for the highest sampling fractions. This suggests that, at high sampling fractions, increased bias due to sampling without replacement is counteracted by other factors (such as a larger sample size which usually results in a more diverse set of recruits). Note that bias is essentially zero when the target group has the same mean degree as the rest of the network, so many variables will exhibit minimal bias regardless of sampling fraction. As discussed above, approximately 85% of variables in RDS public health studies have degree ratios ≤ 1.5 and will therefore exhibit bias no greater than five percentage points, and the remaining 15% may exhibit more extreme levels of bias. Researchers should note the potential for more significant amounts of bias when extreme degree ratios are observed in an RDS study.

Both overall bias and sampling variability for samples drawn without replacement are essentially identical to the respective quantities for samples drawn with replacement for sampling fractions up to 20%. The implication of this result is that for sampling fractions of up to 20%, violations of the sampling-with-replacement assumption inherent in the Heckathorn 2007 RDS estimator can be considered negligible. This range includes RDS studies conducted in large cities, such as the CDC NHBS study of IDUs in 23 large US cities noted above, where the median sampling fraction was 2.3%, and studies of jazz musicians in New York City and San Francisco, with a maximum sampling fraction between the cities of 1.6%. For sampling fractions in the 20% to 40% range, the sampling-with-replacement assumption is a modest source of bias, with a magnitude of no more than two percentage points across the range of simulated parameters. Such cases may arise when RDS is employed in small towns, as in the case from Connecticut described above. Here, results should be interpreted with the potential for small amounts of bias in mind, especially for variables that have a strong relationship with degree.

The simulated 95% confidence interval is nonlinear for the branching cases drawn without replacement. The 95% confidence intervals for these cases shrink for sampling fractions of up to 40% and then diverge, so at very high sampling fractions, the RDS estimates for without-replacement branching cases for target groups with lower and equal mean degree have very wide 95% confidence intervals.

Our results help map the bias and recruitment failure spaces of the Heckathorn (2007) RDS estimates. Just as previous work (Heckathorn 2010) mapped design effect space, so we describe a parameter space where the Heckathorn (2007) RDS estimates produce different levels of bias. The bias is nearly constant in some parameters (homophily) and highly nonlinear in others (sampling fraction), with minimal bias in the parameter ranges of homophily $<.7$, and sampling fraction below 20%.

Our work has two simple implications for empirical RDS studies employing the Heckathorn (2007) estimator: in order to minimize bias, keep sampling fraction below 20%, and avoid very high ($>.7$) homophily networks. Our study points to the importance of presurvey ethnographic and field research to detect high levels of homophily and the use of empirical research such as capture-recapture methods to calculate the sampling fraction. We show that for small sampling fractions, RDS produces low levels of bias, but we do not rule out the possibility of using alternative estimators for very large sampling fractions.

6. Limitations

The principal goal of this article is to examine the effect of violation of the sampling-with-replacement assumption of RDS on bias of RDS population proportion estimates, specifically the effect of this violation for studies with high sampling fractions. The methodology of this article is a mix of analytic and simulation approaches. In this article, we have chosen to focus on our goal and methodology rather than explore many possible implications for RDS bias (and many possible methods); as a result, our analysis has several limitations.

First, we *only* explore the violation of one RDS assumption—sampling without replacement—and not the violation of other assumptions. Specifically, we do not explore violations of the assumption of random recruitment, which we repeat here:

Assumption 3.5 Respondents recruit as though they are selecting randomly from their neighborhoods.

This assumption states only that respondents recruit as though they are selecting randomly—not that respondents employ a truly random process when choosing recruits. There are complicated or nonsystematic ways that may achieve unbiased results without being truly random. At the same time, this assumption indicates that studies where recruitment has a particular bias (e.g., all respondents recruit people of the same gender) will produce biased estimates. This assumption is certainly worth examining, but is far beyond the scope of this article. We are interested in exploring this assumption in future work.

Second, our analysis focuses on targets of recruitment, not sources. We do not consider, for example, whether a respondent who tries to make multiple recruitments but fails (because they have already been recruited) will get discouraged and stop trying to recruit others. These questions are certainly worth investigating, but they are less analytically tractable than the issue of multiple recruitment attempts. For example, we have derived results about multiple recruitment attempts from assumptions about the underlying network structure and sampling fraction. Common network structures have been studied extensively in the literature analyzing social networks. Sampling fractions have been determined in meta-analyses of RDS studies (Wejnert et al. 2012). In contrast, in order to derive results about respondent behavior we would have to start with assumptions about the enthusiasm levels and so on of participants in RDS studies, which to our knowledge have not been explored in the RDS literature. Consequently, we think that a follow-up, empirical study of RDS participants would be better suited to analyzing RDS assumptions with respect to recruiter behavior.

Third, our analytic and simulation results examine the bias in the estimate of population proportions and not the sampling variance of the estimate. In our simulations, the difference between with- and without-replacement sampling variance increases with sample size and with-replacement sampling has lower sampling variance. This contrasts with Lu et al. (2012) who found less sampling variance in simulated without-replacement sampling on an empirical online social network. These different findings suggest that the relationship between sampling fraction and sampling variance may depend on currently unidentified elements of network structure. Sampling variance is clearly an important consideration that should be addressed in future work.

Fourth, we do not analytically link the proportion of repeated sampling events to the magnitude of the bias. The proportion of repeated sampling events is a way to quantify the divergence between with- and without-replacement sampling. We demonstrate both analytically and in simulations (in Figures 1–3) that the density of repeated sampling events (proxied by the proportion of failed recruitments in the simulation section) increases smoothly with the sampling fraction. The important analytic result is that this quantity grows in a bounded and predictable way as sampling fraction increases. Unfortunately, there are many possible sources of bias in RDS studies, including recruiter activity, recruiter preference for certain recruits, and failed recruitments. Most of these sources of bias are not analytically tractable. We are able to demonstrate via simulation that the increase in the proportion of repeated sampling events is associated with an increase in the bias from sampling with replacement. Our simulation results demonstrate a clear positive and near-linear relationship between the proportion of repeated sampling events and sampling fraction and a clear positive relationship between sampling fraction and the magnitude of estimate bias when the subgroups in the target population have different mean degree.

Finally, we do not study empirical social networks of RDS participants. Such networks are extremely hard to collect: most RDS studies ask participants about the number, but not the exact identity, of their friends. This lack of specificity is crucial for privacy reasons; at the same time, it leads to a lack of knowledge about the empirical networks surveyed via RDS. There is no prima facie reason to assume the networks typically targeted by RDS surveys are structured in a fundamentally different way than fully mapped social networks; at the same time, empirical networks may have unusual features that lead to bias in RDS studies. For example, one individual may serve as the only broker between two otherwise physically and socially separated groups, such as one dealer connecting two groups of drug users in different neighborhoods. Such “chokepoints” would be extremely problematic for RDS studies, and yet for the abovementioned reason there are no empirical studies to our knowledge that investigate the frequency of these network structures in RDS studies. We hope that in future work we can study the question of whether additional, noninvasive questions during RDS studies can help researchers identify chokepoints or other unusual network structures in the field without violating participant privacy. Ultimately, we would like to come up with recommendations for researchers to dynamically adjust their sampling strategy when they encounter an unusual network structure, so as to minimize bias in the resulting RDS sample.

7. Conclusion

Our analysis has described a large parameter space of possible conditions for RDS studies, and the levels of bias across this space. We have shown that for a wide range of parameter values, mean bias remains extremely low, and the biased estimate is not significantly different from the true value under conditions which reasonably correspond to empirical RDS studies. We have also shown that higher levels of bias due to sampling without replacement do not necessarily correspond to higher levels of overall bias; on average, sampling without replacement is neither the only nor the dominant factor affecting RDS estimates.

Our results suggest that bias is negligible for sampling fractions up to 20%, a case which fits most studies of hidden populations in large urban settings, for example, the abovementioned studies where sampling fractions range from 0.6% to 8%. In the 20–40% range, the magnitude of bias depends on other parameters, especially whether the variable of interest is correlated with degree. Bias may be as much as two percent if the degree ratio is 1.5; if the variable is independent of degree, bias is again negligible. This case fits studies in small towns or sparsely populated rural areas, or studies in large cities with very large sample sizes. In the 40% to 80% range, the biases of the RDS estimator and the raw sample proportion tend to be in opposite directions, so an estimate in between these two will be less biased. Gile’s (2011) successive sampling approach is a principled method to mediate between the RDS estimator and the raw sample proportion as a function of sampling fraction. Finally, at very high sampling fractions, the sample is best treated as a census rather than a sample, so no statistical estimation process is required. This fits studies conducted either in very sparsely populated areas, or studies that are sufficient to saturate the target population.

An implication of these suggestions is that population-size estimation should be incorporated into all RDS studies; otherwise the most appropriate mode of analysis cannot be identified. A further implication is that population estimation should involve quantitative procedures such as capture-recapture or network scale-up, because estimates even from knowledgeable key informants can be wrong by more than an order of magnitude (Heckathorn et al. 2002).

This article introduced a new theoretical concept for analyzing bias in RDS analyses, a *repeated sampling event*, in which a peer-recruitment attempt fails because the respondent has already participated in the study. Analysis of the density of these events provides a new conceptual tool for analyzing bias in RDS analysis. It also has implications for research design. For example, in sparse networks where branching is precluded (i.e., respondents can recruit only a single peer), recruitment chains tend to die out quickly, so attaining a desired sample size may involve employing very large number of seeds; in extreme cases, more than half the sample may be seeds. Because RDS seeds contribute less information to the sample than a peer-recruited participant but cost the survey the same in terms of time and participation incentives, the efficiency of the project may be severely compromised. Furthermore, if a significant proportion of the sample is composed of seeds, RDS weights are not appropriate for calculating point estimates. However, if branching is permitted (e.g., allowing each respondent to recruit three peers), the number of seeds required to attain a specific sample size is dramatically reduced, thereby increasing the efficiency of the study. Hence, though we confirmed previous findings that branching can increase a study's design effects under some conditions (Goel and Salganik 2009), or have trivial effects under others (Heckathorn 2002), we also identify a compensatory benefit from building branching into a research design: in sparse networks, it can greatly increase a study's efficiency.

A final implication of the study is to demonstrate the importance of exploring a large parameter space when quantifying bias in RDS studies, for studies which explore only limited regions may produce misleading results, especially if the region investigated fails to encompass the full range of RDS studies reported in the literature. We explore sampling fractions between five percent and 80%, homophily between 0 and .7, and various network structures and degree distributions to produce results that can provide practical insight about the impact of nonreplacement bias on RDS estimates.

The supplemental data is available at: www.dx.doi.org/10.1515/JOS-2016-0002

Appendix: Relating This Analysis to RDS Work Previously Published in JOS

Implications of Violation of Sampling With Replacement for Volz-Heckathorn Estimator.

The Volz-Heckathorn estimator relies on a model of chain-referral sample as a random walk on a network. In the case of sampling with replacement, this model is accurate, and the random walk is a Markov Process, which in equilibrium occupies a node with probability proportional to degree (Salganik and Heckathorn 2004). However, in the case of sampling without replacement, the model is inaccurate: instead of being a random walk on a network, the RDS sample is a self-avoiding walk (SAW) on a finite network. Since the network is finite and the SAW may not by definition visit a node more than once, in equilibrium the probability of it occupying any node approaches 0. In this case, we can no longer apply the arguments in Salganik and Heckathorn (2004), but must propose another model of RDS as a self-avoiding random walk.

By definition, an RDS sample is a finite-size chain-referral sample, in which no individual may be recruited more than once, drawn from a larger (but still finite) network. Thus, we can formalize RDS as a length $|R|$ SAW on a network of size $|P|$. This SAW corresponds to a reducible Markov Process MP^{WOR} on the set of nodes in R, where each state is a node and transitions between states correspond to recruitments. Note that for this reducible Markov Process, no state may have more than one incoming transition from another state. The reducible Markov Process has a number of differences to the irreducible Markov Process MP^{WR}, which models sampling with replacement. However, both processes may be encoded as transition matrices, and we can compare the transition matrices to measure the extent of bias due to sampling with replacement.

We can construct an incidence matrix M for the network, where for any individuals *i, j* in the larger population P, $R_{ij} = 1$ if i and j are connected, 0 otherwise. This matrix gives the equilibrium transition probabilities for MP^{WR}. Indeed, we can construct a *transition matrix* M^{WR} where for any individuals *i, j* in P, $M^{WR}_{ij} = 1$ if i recruited j into the chain-referral sample with replacement modeled by MP^{WR}, 0 otherwise. This matrix will approximate M in the sense that, for some node i, the larger *i's* in-degree ($\sum_i M_{ij}$), the more likely and more frequently will i be recruited in the chain-referral sample ($\sum_i M^{WR}_{ij}$). Similarly, we can construct a transition matrix M^{WOR} where for any individuals *i, j* in *P*, $M^{WOR}_{ij} = 1$ if i recruits j to participate in the RDS study, and 0 otherwise.

Note that if no individual is ever recruited more than once in the course of the chain-referral process modeled by MP^{WR}, then $M^{WOR} = M^{WR}$. However, even a single repeated sampling event can introduce a chain of differences between the two transition matrices—for example, let A be a with-replacement sample wherein i recruits j who recruits k who recruits i who recruits l. Let B be a repeated sampling sample wherein i recruits j who recruits k, and then the sample ends. The corresponding transition matrices would look as follows:

A

	i	j	k	l
i	0	1	0	1
j	0	0	1	0
k	1	0	0	0
l	0	0	0	0

B

	i	j	k	l
i	0	1	0	0
j	0	0	1	0
k	0	0	0	0
l	0	0	0	0

Still, by definition any atomic (cell-level) differences between M^{WOR} and M^{WR} are due entirely to repeated sampling events, and so we can operationalize the bias due to sampling with replacement as the difference between the two matrices.

8. References

Barabási, A.L. and R. Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286: 509–512. Doi: http://dx.doi.org/10.1126/science.286.5439.509.

Bernard, H.R., T. Hallett, A. Iovita, E.C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M.J. Salganik, T. Saliuk, O. Scutelniciuc, G.A. Shelley, P. Sirinirund, S. Weir, and D.F. Stroup. 2010. “Counting Hard-to-Count Populations: the Network Scale-Up Method for Public Health.” *Sexually Transmitted Infections* 86 (suppl. II): ii11–ii15. Doi: http://dx.doi.org/10.1136/sti.2010.044446.

Bernhardt, A., M.W. Spiller, and N. Theodore. 2012. “Employers Gone Rogue: Explaining Industry Variation in Violations of Workplace Laws.” *Industrial and Labor Relations Review*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2013376 (accessed January 2016).

Curtis, R., K. Terry, M. Dank, K. Dombrowski, and B. Khan. 2008. *The Commercial Sexual Exploitation of Children in New York City, Volume One: The CSEC Population in New York City: Size, Characteristics, and Needs*, Final report submitted to the National Institute of Justice. New York: Center for Court Innovation and John Jay College of Criminal Justice. Available at: https://www.ncjrs.gov/pdffiles1/nij/grants/225083.pdf (accessed January 2016).

Erdős, P. and A. Rényi. 1959. “On Random Graphs.” *Publ. Math* 6: 290–297.

Gile, K.J. 2011. “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation.” *Journal of the American Statistical Association* 106: 135–146. Doi: http://dx.doi.org/10.1198/jasa.2011.ap09475.

Gile, K.J. and M.S. Handcock. 2010. “Respondent-Driven Sampling: An Assessment of Current Methodology.” *Sociological Methodology* 40: 285–327. Doi: http://dx.doi.org/10.1111/j.1467-9531.2010.01223.x.

Goel, S. and M.J. Salganik. 2009. “Respondent-Driven Sampling as Markov Chain Monte Carlo.” *Statistics in Medicine* 28: 2202–2229. Doi: http://dx.doi.org/10.1002/sim.3613.

Hagberg, A.A., D.A. Schult, and P.J. Swart. 2008. “Exploring Network Structure, Dynamics, and Function Using NetworkX.” In Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA, August 2008. Edited by G. Varoquaux, T. Vaught, and J. Millman. 11–15.

Heckathorn, D. 1997. “Respondent-Driven Sampling: a New Approach to the Study of Hidden Populations.” *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Heckathorn, D.D. 2002. “Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations.” *Social Problems* 49: 11–34. Doi: http://dx.doi.org/10.1525/sp. 2002.49.1.11.

Heckathorn, D.D. 2007. “Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment.” *Sociological Methodology* 37: 151–207. Doi: http://dx.doi.org/10.1111/j.1467-9531.2007.00188.x.

Heckathorn, D.D. 2010. “Sampling Elusive and Hard-to-Reach Populations: The Scope and Limits of Respondent-Driven Sampling.” Presented at the Duke Population Research Institute Workshop “Challenging Samples: Networks and Surveys in Demographic and Health Research” on May 7, 2010.

Heckathorn, D.D. 2011. “Snow ball versus Respondent-driven Samping.” *Sociological Methodology* 41: 355–366. Doi: http://dx.doi.org/10.1111/j.1467-9531.2011.01244.x.

Heckathorn, D.D. and J. Jeffri. 2001. “Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians.” *Poetics* 28: 307–329. Doi: http://dx.doi.org/10.1016/S0304-422X(01)80006-1.

Heckathorn, D.D., R.S. Broadhead, and B. Sergeyev. 2001. “A Methodology for Reducing Respondent Duplication and Impersonation in Samples of Hidden Populations.” *Journal of Drug Issues* 31: 543–564.

Heckathorn, D.D., S. Semaan, R.S. Broadhead, and J.J. Hughes. 2002. “Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Ages 18–25.” *AIDS and Behavior* 6: 55–67. Doi: http://dx.doi.org/10.1023/A:1014528612685.

Iguchi, M.Y., A.J. Ober, S.H. Berry, T. Fain, D.D. Heckathorn, P.M. Gorbach, R. Heimer, A. Kozlov, L.J. Ouellet, S. Shoptaw, and W.A. Zule. 2009. “Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications.” *Journal of Urban Health* 86 (Suppl. 1): 5–31. Doi: http://dx.doi.org/10.1007/s11524-009-9365-4.

Jeffri, J., D.D. Heckathorn, and M.W. Spiller. 2011. “Painting Your Life: a Study of Aging Visual Artists in New York City.” *Poetics* 39: 19–43. Doi: http://dx.doi.org/10.1016/j.poetic.2010.11.001.

Lansky, A., A. Drake, and H.T. Pham. 2009. *HIV-Associated Behaviors Among Injecting-Drug Users – 23 Cities, United States, May 2005-February 2006*. Morbidity and Mortality Weekly Report April 10, 2009, 58: 329–332. Available at: www.cdc.gov/mmwr/pdf/wk/mm5813.pdf (accessed January 2016).

Lee, C.-H., X. Xu, and D.Y. Eun. 2012. “Beyond Random Walk and Metropolis-Hastings Samplers.” In Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems – SIGMETRICS’12, 319. ACM Press. Doi: http://dx.doi.org/10.1145/2254756.2254795.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. “The Sensitivity of Respondent-Driven Sampling.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 175: 191–216. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

Malekinejad, M., L.G. Johnston, C. Kendall, L.R.F.S. Kerr, M.R. Rifkin, and G.W. Rutherford. 2008. “Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review.” *AIDS and Behavior* 12(suppl. 4): 105–130. Doi: http://dx.doi.org/10.1007/s10461-088-9421-1.

Maslov, S. and K. Sneppen. 2002. “Specificity and Stability in Topology of Protein Networks.” *Science* 296: 910–913. Doi: http://dx.doi.org/10.1126/science.1065103.

Ramirez-Valles, J., D.D. Heckathorn, R. Vázquez, R.M. Diaz, and R.T. Campbell. 2005. “From networks to populations: the development and application of respondent-driven

Sampling Among IDUs and Latino Gay Men." *AIDS and Behavior* 9(4): 387–402. Doi: http://dx.doi.org/10.1007/s10461-005-9012-3.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

Smylie, J., M. Firestone, L. Cochran, C. Prince, S. Maracle, M. Morley, S. Mayo, M.W. Spiller, and B. McPherson. 2011. *Our Health Counts Urban Aboriginal Health Database Research Project – Community Report First Nations Adults and Children, City of Hamilton*. Hamilton: De Dwa Da Dehs Neys Aboriginal Health Centre. Available from: http://www.stmichaelshospital.com/crich/wp-content/uploads/ourhealth-counts-report.pdf (accessed January 2016).

Tillé, Y. 2006. *Sampling Algorithms*. New York: Springer.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24: 79–97.

Wallace, R. 1991. "Social Disintegration and the Spread of AIDS: Thresholds for Propagation Along 'Sociogeographic' Networks." *Social Science & Medicine* 33: 1155–1162. Doi: http://dx.doi.org/10.1016/0277-9536(91)90231-Z.

Wang, J., R.G. Carlson, R.S. Falck, H.A. Siegal, A. Rahman, and L. Li. 2005. "Respondent-Driven Sampling to Recruit MDMA Users: A Methodological Assessment." *Drug and Alcohol Dependence* 78: 147–157. Doi: http://dx.doi.org/10.1016/j.drugalcdep. 2004.10.011.

Watts, D.J. and S.H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393(6684): 440–442. Doi: http://dx.doi.org/10.1038/30918.

Wejnert, C., H. Pham, N. Krishna, B. Le, and E. DiNenno. 2012. "Estimating Design Effect and Calculating Sample Size for Respondent-Driven Sampling Studies of Injection Drug Users in the United States." *AIDS and Behavior* 16(4): 797–806. Doi: http://dx.doi.org/10.1007/s10461-012-0147-8.

Received August 2014
Revised July 2015
Accepted September 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 75–92, http://dx.doi.org/10.1515/JOS-2016-0003

Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes

Frederick G. Conrad[1], *Mick P. Couper*[1], *and Joseph W. Sakshaug*[2]

A source of survey processing error that has received insufficient study to date is the misclassification of open-ended responses. We report on efforts to understand the misclassification of open occupation descriptions in the Current Population Survey (CPS). We analyzed double-coded CPS descriptions to identify which features vary with intercoder reliability. One factor strongly related to reliability was the length of the occupation description: longer descriptions were less reliably coded than shorter ones. This effect was stronger for particular occupation terms. We then carried out an experiment to examine the joint effects of description length and classification "difficulty" of particular occupation terms. For easy occupation terms longer descriptions were less reliably coded, but for difficult occupation terms longer descriptions were slightly more reliably coded than short descriptions. Finally, we observed as coders provided verbal reports on their decision making. One practice, evident in coders' verbal reports, is their use of informal coding rules based on superficial features of the description. Such rules are likely to promote reliability, though not necessarily validity, of coding. To the extent that coders use informal rules for long descriptions involving difficult terms, this could help explain the observed relationship between description length and difficulty of coding particular terms.

Key words: Survey processing error; coding error; occupational classification.

1. Introduction

Survey responses are imperfect measures. The origins and implications of response error are increasingly well understood (e.g., Sudman et al. 1996, ch. 2; Tourangeau et al. 2000, ch. 1) but the vast majority of this knowledge concerns closed-form questions, i.e., questions that present response options which respondents select to report their answers. However, essential information for official statistics is derived from open responses, i.e., answers reported in respondents' own words. These open responses are coded – assigned to categories – in order to be quantified, and the coding process can introduce

[1] Survey Research Center, Institute for Social Research, University of Michigan, PO Box 1248, Ann Arbor, MI 48106, U.S.A., and Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, MD 20742, U.S.A. Emails: fconrad@umich.edu and mcouper@umich.edu
[2] Cathie Marsh Institute for Social Research, University of Manchester, Humanities Bridgeford Street Building, Manchester M13 9PL, U.K., and Institute for Employment Research, Regensburger Str. 104, Nuremberg 90478, Germany. Email: joe.sakshaug@manchester.ac.uk

Acknowledgments: We wish to acknowledge the assistance and advice of Adelle Belinger, Laverne Bowman, Cathy Dippo, Janet Harmon, Fred McKee, George Mitcham, Ted Sands, Ron Tucker, John Van Hoewyk, and Greg Weyland. We are particularly indebted to the coding staff at the Census Bureau's Data Processing Center for its participation in the study reported here. We thank Lars Lyberg for making helpful suggestions about an earlier version of this article. Finally, we are grateful to four anonymous reviewers and the Associate Editor for their valuable comments that improved the article.

error to the survey data. Coding error does not necessarily mean that a code is "inaccurate." Responses can vary in the degree to which they belong in particular coding categories, i.e., category membership is graded (e.g., Barsalou 1985), so simply assigning a code to an open response can involve error because the fit is not perfect, even if there is no better fitting code.

One domain in which open responses are essential is the measurement of occupation. The coding of occupation has many unique features compared to coding other types of open responses: coders must assign a description to one of hundreds of possible codes rather than just a handful, the open answers are short and factual rather than long and attitudinal, and occupation coders who make this their profession become very skilled. The data about occupation that are produced through the coding process are commonly used to study various phenomena occurring in the labor force, including sex segregation (Anker 1998), work-related injuries (Cawley and Homce 2003; Layne 2004; Reichard and Jackson 2010), health-related exposures (Kromhout et al. 1993; Hammond et al. 1995; Kauppinen et al. 2000), wage inequality (Lettau 2003; Heywood and O'Halloran 2005; Bjerk 2007), mobility (Shniper 2005; Moscarini and Thomsson 2007), and occupational projections (Rosenthal 1992).

Despite differences in the details of coding open responses in different domains, there seem to be certain commonalities in the ways coders contribute to overall error. Coder error is a type of processing error (Biemer and Lyberg 2003, 234–241) that is introduced by coders interpreting respondents' verbalization of their thinking. By some measures, it can substantially inflate other sources of error. For example, in a study of time use, Sturgis (2004) demonstrated that correlated coder error – analogous to interviewer variance – nearly doubled the size of standard errors on average, across ten activity categories with a maximum inflation factor of 3.36. Classification of open reports may be compromised because, for example, coders may fail to consider key information such as the size of different categories and, thus, the probability of membership. Base rates are just one consideration when all else is equal or there is no way to choose between alternative classifications. Alternatively, the quality of coding may be degraded because the coding categories and the rules for using them are too rigid to adequately address the ambiguity inherent in people's descriptions; for example, respondents' descriptions may fit well into more than one category or may not fit well into any category but must be assigned to one nevertheless.

There are various techniques for coding open-ended responses into different categories. The majority of these techniques can be classified into three distinct groups: manual coding, computer-assisted coding, and automated coding. In manual coding, respondents provide their response in free text and coders assign codes based on a standardized classification system. In computer-assisted coding, coders provide codes by means of assistance dictionaries, which provide lists of possible answers for the coder to choose from (Bushnell 1995; Lyberg and Kasprzyk 1997). In automated coding, the software codes some open-ended responses without human intervention, and manual (or computer-assisted) coding is used to code the unresolved balance (Lyberg and Dean 1992; Macchia and D'Orazio 2001; Esuli and Sebastiani 2010).

Using occupation as an example, in the current work we explore how (1) the characteristics of respondents' open-ended reports – in particular the length of the reports and the difficulty of classifying descriptions that include particular words – and (2) the practices of coders may affect the quality of *computer-assisted* coding of open-ended answers.

The occupation coding task that we study here involves assigning respondents' occupation descriptions, collected from open-ended reports of their occupation and duties, to categories in the US Census occupational classification system, which is derived from the Standard Occupational Classification System (SOC) and has been updated several times. The SOC is one of several standard classification systems established by the US Office of Management and Budget, and is used to publish comparable occupational data for statistical purposes across the U.S. federal statistical agencies. The data set used in this study involved codes from the 1990 Census system, derived from the 1977 SOC. The 1990 Census system included 501 detailed categories, expressed as three-digit codes, 13 major occupation groups, expressed as two-digit codes, and six summary groups, expressed as one-digit codes. There are some differences between the 1977 SOC and the most recent version, the 2010 SOC. The number of major occupation groups has increased from 21 to 23, and the number of detailed occupations has increased from 662 to 840. A detailed description of the revision history and process can be found online (Bureau of Labor Statistics, 2014).

While occupation is one important domain in which the data are based on coded open-ended responses, coding open responses is ubiquitous throughout survey research, producing important data in domains ranging from public opinion (e.g., most serious problems facing the country) to time use (activities) to academic fields of study. To the extent that the coding of open responses across domains and question types is similar to the coding of occupation descriptions, what we learn about occupational coding may inform how we think about coding open responses in other applications.

2. Classification and Measurement

The coding process may introduce error in several ways having to do with (1) the "in or out" requirement of a formal classification system in a world where categories are ever changing and where membership is a matter of degree, (2) lack of category size information to help inform coder decisions when a description seems to fit two categories equally well, (3) ambiguity of particular words that respondents use in occupational descriptions, and (4) length of occupation descriptions.

In everyday classification tasks, people can modify categories to accommodate instances that are atypical. However, formal classification systems are more rigid than this. In a demonstration of the flexibility of everyday categories (as opposed to a formal classification system like the SOC), Kunda and Oleson (1995) found that when presented with descriptions of cases that deviate from the stereotype (e.g., an introverted lawyer), under the right circumstances people preserved the integrity of the main category (lawyer) by creating subtypes (a category for introverted lawyers). Under other circumstances, they redefined the main category to include deviant cases (lawyers in general were rated as more introverted than they were if no deviant case had been presented). A coder using an established classification scheme would not be able to accommodate the deviant case in either of these ways, but instead would have to assign it to a category despite the poor fit.

When respondents' answers are ambiguous, that is, could be assigned to more than one category, coders may systematically assign answers to the wrong category. Tversky and Kahneman (1983) found that if an instance sounds like it could belong to the conjunction of two everyday categories (e.g., a feminist bank teller) but could also be assigned to one of the individual categories (bank teller), people are more likely to judge such instances to be members of the conjunction than the individual category. They call this the *conjunction fallacy* because there are more bank tellers than feminist bank tellers in the world, yet people seem to give more weight to similarity of the description to the category prototype (sounds more like a feminist bank teller than a bank teller) than to the size of the category (there are more bank tellers than feminist bank tellers). In another demonstration of this general tendency, Tversky and Kahneman (1974) found that people were more likely to judge someone who seemed like an engineer to be an engineer than to be a lawyer, even though they were told that, in the experimental scenario, there were more lawyers than engineers. They call this the *representativeness heuristic* and, while it can be a useful guideline in making some classification judgments, it desensitizes people to the base rate or size of categories when making these judgments. Coders may be similarly oblivious to category size and probability of membership when faced with descriptions that sound like particular categories, even if instances of these categories are relatively rare. This is not to say base rates and probability of category membership should be the *only* consideration that informs a coder's decision. If a description could fit equally well into two categories with very distinct meanings – for example, "secretary" could refer to a senior official of an organization or to an office assistant – a coder could be instructed to choose the category for which the odds of membership are greater rather than flipping a coin. There are more office assistants than senior officials in the world so in the absence of any additional information, considering the size of the categories would be a rational – if imperfect – strategy.

By another view, it is not flaws in coder decision making as much as the descriptions themselves that lead to lower-quality codes. Within a particular domain, some terms in respondents' descriptions may be inherently hard to code, for example, they may fit poorly into existing categories or may fit well into multiple categories. Coders may address this by developing specialized rules for classifying descriptions with problematic terms (see, e.g., Hak and Bernts 1996; Martin et al. 1995). While the use of such rules should increase agreement among coders, this could well happen without any increase in the "validity" of codes. It could be the case that a rule leads to incorrect – but consistent – codes on some occasions because it may lack the means to adjust the code on the basis of subtle changes in context. Such rules may actually lower agreement among coders if they are not defined by the group or, for other reasons, not unanimously endorsed. This is particularly likely when rules are not explicitly documented. Furthermore, one rule may conflict with another even though both seem to apply to a particular case; this too might lead to disagreement.

In addition to the inherent difficulty in coding certain terms, the length of respondents' answers may also affect how well they are coded. Couper and Conrad (1996) asked a national sample standard questions about their occupation ("What kind of work do you do, that is, what is your occupation?") and duties ("What are your usual activities or duties at this job?") from the Current Population Survey (CPS), and asked half of the sample an additional question about their job title ("What is your job title?"). When coders were able to consider the extra response in their coding decision, their agreement with each other was lower than when they only had the initial response to work with. Cantor and Esposito (1992) report that coders prefer less information: they asked coders to listen to the interviews and indicate where additional probes would have helped them code the response. The coders virtually never asked for additional probes, suggesting they recognized that longer descriptions are harder to code than shorter ones.

More information could harm coder agreement for much the same reason that in everyday classification people prefer categories at intermediate levels of abstractness – a concept known as the *basic level* (e.g., Rosch et al. 1976). The idea is that categories that are neither too abstract nor too concrete are most useful, for example, "dog" (basic level) versus "Welsh Terrier" (more concrete) or "mammal" (more abstract). Thus description length may be a proxy for level of abstraction: longer descriptions will facilitate coding to the extent that they refer to basic-level jobs but will confuse matters if they describe overly specific categories.

On the other hand, longer descriptions are likely to be more specific than shorter ones – more words probably convey more detail. Perhaps for this reason, longer open responses are often assumed to be of higher quality than shorter ones across a variety of domains (e.g., Andrews 2005; Smyth et al. 2009; Israel 2010). Moreover, according to the 1997 CPS Field Interviewers Manual, as well as the current manual (U.S. Census Bureau, 2013), interviewers are told that

> One-word responses to the question on occupation (for example, clerk, manager, nurse, engineer, teacher) are usually far too general to be coded accurately. Whenever very brief responses are given, probe to obtain a more specific response.

So, one can imagine agreement would be higher for longer, or at least more detailed, descriptions: with more detail, there is less opportunity for two coders to interpret the description differently.

2.1. *Measures of Coding Quality*

Just as in assessing the quality of closed responses, *validity* and *reliability* are generally used to characterize the quality of open responses. However, the notion of validity is not as straightforward when applied to coded open responses as it is with respect to closed responses, at least for facts and behaviors such as one's job title or one's duties at work. Validity of closed responses can be determined, in principle, by comparing the responses to a gold standard such as a set of administrative records; with coded data, validity is typically operationalized as agreement with an expert. (With automated coding, validity is typically defined as matching an open response to text in a reference dictionary that maps text examples to categories, e.g, Macchia and D'Orazio 2001). This has much of the character of an agreement or reliability measure: a valid code matches another code that is treated as the gold standard; if there is not agreement the response is considered not valid. This lacks the potential for objective verifiability that is part of response validity for closed (factual) responses.

Reliability is simpler in concept – agreement between two or more classifications of an open response – but it is less definitive than a validity measure in that two or more coders can agree with each other without necessarily being “correct.” They can both be “wrong”, assuming the correct category is known or is knowable.

3. Current Study

In the current study we focus on characteristics of respondents’ occupation descriptions from the CPS that might affect the quality of codes. As noted above, the CPS descriptions come from one question about occupation, “What kind of work do you do, that is, what is your occupation?” and one about duties, “What are your usual activities or duties at this job?” After filtering out “special-case” occupations for which direct mappings between descriptions and codes are provided and “combined occupations” for which, again, direct mappings between descriptions and categories are provided, coders are instructed to consider both the occupation and duties responses together in assigning a single numeric occupation code to the description (see U.S. Census Bureau 2014). Consider the following example:

OCC – Credit Manager

DUTIES – Directing operation of credit department

In a case like this, the coder is instructed to combine the word “department” from the DUTIES line with the content of the OCC line (“Credit Manager”) and code “Manager, Credit Department.” Much of the instruction concerns direction on how to proceed beyond an impasse. For example, if the occupation and duties lines contain contradictory information, coders are taught to use whichever is more specific. It is our assessment that the training about the actual coding *decision* is not more detailed than instructions of this sort: coders need to be very familiar with the occupation categories and use their judgment about which words are important to consider and which ones are not. In the end, the coding task relies more on coders’ knowledge of the job definitions and their aptitude for determining which parts of the description to consider than on particular training in the coding procedure. Although coders’ expertise in occupational classification is at the center of the classification process, the coders searched electronic indices for occupation categories corresponding to particular terms contained in the description by entering those terms into a computer. The tool did not classify the description for the coder but returned possible categories given the input. It was still the coders’ decision what category best fit the description. Although not in place at the time of the current study, the Census Bureau introduced an autocoder in 2012 that provided coders with suggested classifications for particular descriptions.

The study we report has three parts. The first is designed to explore what characteristics of occupation descriptions reduce coding reliability. We analyzed twelve months of CPS occupation descriptions (March 1997 to February 1998); note that although these data were collected and coded many years before the current article was written, the Census Bureau confirmed that they currently process and code open-ended responses in essentially the same way they did in 1997–1998. These descriptions represent 32,362 cases, each of which was independently classified by two coders. More specifically, about ten percent of all industry and occupation (I&O) descriptions were double coded, that is, independently classified by a second coder. This process was conducted for quality assurance (QA), not production purposes; that is, the original code was not affected by the second code. Once the second (QA) coder assigned a code, the initial (production) code was revealed, and the coder had to decide whether to change his or her code to the original code (assuming a discrepancy) or refer the case back to the field for more information. The first part of the study consisted of analyses of this data set, including the effects of the length of descriptions.

Second, we investigated how characteristics identified in the first part of the study jointly affect coding agreement. To do this, we created a data set of occupation descriptions that varied systematically on several characteristics and asked pairs of coders to classify them. The experiment explicitly tested the joint effect of coding difficulty of particular words in the description and the length of descriptions. To do so, we created a set of 800 occupation descriptions systematically varied on the following dimensions:

(1) Length: one, two, and three or more words
(2) Difficulty of "primary" word: easy versus hard
(3) Difficulty of "secondary" word: easy versus hard
(4) Order of primary word: first, not first.

The easy primary words were selected by taking the eight words from the QA dataset with the highest agreement ratio. The eight words chosen were: secretary, cashier, driver, cook, teacher, nurse, waitress, and carpenter. A similar process was used to select the hard primary words (high ratio of disagreement to agreement), resulting in the following selection: owner, operator, laborer, director, technician, clerk, supervisor, and administrator. The secondary words were chosen using similar procedures, that is, high ratio of agreement to disagreement and vice versa, conditioning on each of the eight easy and eight difficult primary words first. This produced equal numbers of easy-easy, easy-hard, hard-easy, and hard-hard word pairs, for example, "school nurse" would be an easy-easy word pair. We then randomly selected existing descriptions from the QA data containing these word pairs. While the QA data set contained a large number of descriptions (over 30,000), there were some word pairs for which no description existed in the data set. In these cases, we created new descriptions by adding or removing words from descriptions that partially matched the word pair. For example, if the word pair "research supervisor" was not found in a description but "laboratory supervisor" was, we used that description, including the duties, but substituted "research" for "laboratory."

These 800 descriptions were then seeded into the ongoing production coding process, using the same procedures as regular CPS coding, but with all of the experimental cases being flagged for QA coding. In this way we obtained two codes for each of the experimental descriptions from coders who were blinded to which cases came from the experimental corpus.

Finally, we examined the coders' strategies and the kind of information they brought to bear while performing the coding task. In this third part of the study, we asked coders to think aloud while classifying occupation descriptions excerpted from the set created for the second part of the study. More specifically, we selected 100 cases from the experimental corpus, and observed four coders each coding 50 cases. Multiple-word descriptions from the experiment just described were overselected as these tend to produce higher levels of overall disagreement. The authors interacted with the coders while they were coding, asking them to think out loud about their decision-making process, and probing for reasons for specific actions, roughly following the procedure outlined by Ericsson and Simon (1993).

4. Results

The analyses are reported separately for each of the three parts of the study. In the first part, which concerned coder agreement in the QA dataset, we first report descriptive statistics about agreement and disagreement, then examine whether disagreement is concentrated at certain digits in the occupation codes. This is followed by analyses of disagreement by occupation category. Finally we examine how attributes of the description, in particular the length of the description, affects agreement. Although the second coders could change their classifications once the first coders' classification was revealed, our focus here is on the initial code assigned to each case by the two coders, a cleaner measure of agreement. While the data we examined included both industry and occupation descriptions, we only analyze agreement on occupation, that is, not industry classification.

In the second part (coding experiment) we test whether any length effects observed in the first part are replicated in the experiment. We also test whether length interacts with the "difficulty" (agreement to disagreement ratio) of words in the descriptions. In the third part (coder observation) we analyze the coders' verbal reports, in particular, monitoring for evidence of what knowledge and conventions they use to facilitate coding of ambiguous cases.

4.1. *Analysis of Agreement in Quality Assurance Data*

Table 1 contains more details on the outcome of the double-coding process. A referral implies that the coder has insufficient information to classify the case, and refers the case back to the field for more information. Our main focus is on the 2,749 occupation descriptions (8.5% of all double-coded descriptions) where both coders assigned a code

Table 1. Occupation code agreements, referrals and disagreements

Outcome	Number	Percent of all cases	Percent of nonreferred disagreements
Agreements:	27,518	85.0	
Agreement on substantive code	23,116	71.4	
Agreement on referral	4,402	13.6	
One coder refers	2,095	6.5	
Disagreements:	2,749	8.5	100.0
Disagreement on first digit	1,251	3.9	45.5
Disagreement on second digit	888	2.7	32.3
Disagreement on third digit	610	1.9	22.2
Total	32,362	100.0	

but they disagree. In 22.2% of these cases ($n = 610$), the disagreements were relatively trivial, involving only the last of three digits in the code. However, for the balance of cases the disagreements are more severe, with 45.5% ($n = 1{,}251$) involving the first digit of the code, and a further 32.3% ($n = 888$) involving the second digit. In other words, 3.9% of all cases yield genuine and major substantive disagreements between pairs of coders. As indicated before, agreement does not guarantee accuracy – both coders could be wrong – but disagreement guarantees at least one coder is wrong – both cannot be right.

Given that the coders are evaluated on both speed and accuracy, we suspected that some of the errors may be due to "slips" (e.g., Norman 1981) such as transpositions (e.g., 234 versus 243) or single-digit offsets (e.g., 123 versus 223). We found that simple transposition errors account for a very small fraction (0.2%) of discrepancies. While one-digit offsets account for almost seven percent of discrepancies, many of these may be intended: at the first digit, the substantive difference between categories is large, for example, *legal* vs. *health care*. The majority of the descriptions whose code differed by one digit (247 out of 318) involved a discrepancy on the last digit. In the occupation coding system this represents a minor distinction in the detailed coding scheme, for example, between bartenders (code 434) and waiters/waitresses (code 435). We concluded that slips of this sort are a negligible source of error in occupation coding.

Another issue we explored was whether disagreements were more likely to occur between certain occupation groups. Restricting our focus to those cases where both coders assigned a substantive code and disagreed on the summary group (i.e., the first digit), we found that 29.4% of all these disagreements occurred between two summary groups: (1) managerial and professional specialty occupations, and (2) technical, sales and administrative support occupations. A further 14.8% occurred between (5) precision, production, craft and repair occupations and (6) operators, fabricators and laborers. However, groups (1) and (2) account for only 11% of all occupation codes, while (5) and (6) account for 4.7%. So while there appears to be some clustering of disagreements, the majority of disagreements occur between all summary (first-digit) occupation groups.

While some job categories may be particularly prone to disagreement, the descriptions themselves may affect agreement. One attribute of the descriptions that is potentially relevant to coding agreement is their length. Figure 1 shows the relationship between the number of words in the occupation description (the combined responses to both the occupation and duties questions) and the percent of all cases that result in disagreements

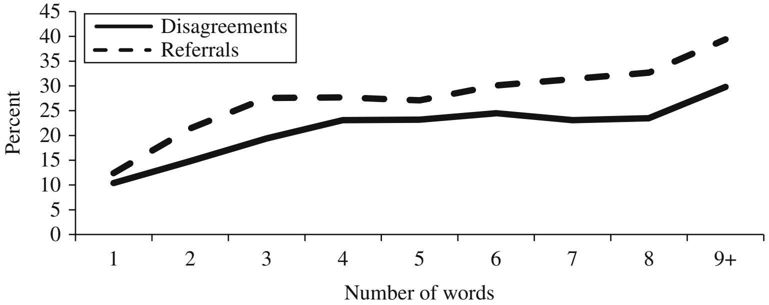


Fig. 1. Length of occupation description and disagreement and referral rates; percent is out of 32,362 cases

(including any disagreements, irrespective of the digit at which the disagreement takes place) and referrals (where at least one coder referred the case back to the field) respectively.

It is clear from this graph that the disagreement rate increases with increasing length of the occupation description. Another way to see this relationship is to compare the mean number of characters and words for the agreements and disagreements. The mean number of characters in descriptions where coders agreed is 15.3 (s.d. 8.69) compared to a mean of 18.6 (s.d. 9.92) for disagreements, a statistically significant ($t = 23.82$; $p < .001$) difference. Similarly, the mean number of words is 2.10 (s.d. 1.20) for agreement cases and 2.56 (s.d. 1.52) for disagreement cases, which is also a statistically significant difference ($t = 20.56$; $p < 0.001$). Given that referrals to the field simply requested *more* information (although coders now indicate *what* information they need), these results suggest that such an approach may actually have been counterproductive. One explanation is that more words simply create more opportunity for coders to disagree; each word is open to interpretation, and given the variability in how different people interpret the same words, the longer the description the greater the chance of disagreement. Of course there are situations in which more information can help clarify a description, for example, if the response is "teacher" and there are only three possible codes "preschool and kindergarten teacher," "elementary and middle school teacher," and "secondary school teacher," clearly more information could disambiguate the description. But if descriptions are already appropriately detailed then more information – more words – can muddle the picture unless they correspond exactly to the definition. A one-word description may well be too abstract (above the basic level) to be reliably classified, so additional words may help. However, as more words are provided – unless they are very similar to the actual definition of the job category – they are likely to confuse coders.

These results concur with Couper and Conrad's (1996) findings mentioned at the outset: they administered the standard CPS occupation and duties questions and then asked half of the sample an additional question about job title. The first-digit coder agreement rate for the standard CPS questions was 86.4% while that for the group asked the additional question was 82.1%. One explanation offered for this finding was that the addition of the job title question reduced the amount of information provided in the occupation description (combined responses of occupation and duties). In fact, the opposite occurred; when job title was asked before the occupation and duties questions, the occupation description was significantly longer than when job title was not asked (23.5 versus 18.0 characters). Furthermore, the length of the occupation description was negatively associated with coder agreement. For example, the average length of the occupation description was 20.5 characters when the coders disagreed on the first digit of the code, but 17.5 when they agreed on the code.

Similar results are presented in an unpublished report (Westat/AIR 1989): The coder agreement rate on summary (first-digit) occupation group using the standard CPS questions was 88%, but only 75% when additional job identification questions were asked. More specifically, the study compared agreement when the standard CPS questions (occupation and duties) were asked to agreement after two additional questions, including one job title probe about the identity of the respondent's job ("What was . . .'s job at [organization name]? *If necessary, probe:* What was . . .'s job title at [organization name]?") were asked. While it is not clear from the report how often the probe was

Table 2. Words with high ratios of disagreement to agreement and agreement to disagreement (ratios in parentheses)

High disagreement to agreement ratios:
Administrative (3.16); services (2.76); research (2.63), assist (2.34); maintenance (2.16); administrator (2.15); general (2.11); service (2.03)
High agreement to disagreement ratios:
Waitress (18.54); registered (8.24); guard (6.45); carpenter (6.34); electrician (5.24); secretary (5.19); accountant (5.16)

actually administered in this experimental condition, the effect of these extra questions could only have been to increase the length of the description and amount of information compared to the standard CPS approach. This again lends support to the finding that the provision of additional information (either in longer descriptions or through additional questions or probes) is associated with lower levels of coder agreement. These consistent findings that appear to run counter to common practice (seeking more information in the case of uncertainty or disagreement) are certainly worth further exploration.

To elaborate further on the length effect, we speculated that length may interact with particular occupations (or words used to describe them) in affecting agreement. For example, some occupations may be easily described using a single word (e.g., waiter), and adding words to the description may only serve to muddy the decision. On the other hand, some occupations may be inherently complex, and cannot be described adequately using only one or two words.

To test this, we separated the QA data file into agreement cases and disagreement cases, and then measured the frequency of the words in each set of descriptions. We found that words such as "administrative" were three times more likely to appear among the disagreement cases than the agreement cases (a ratio of 3.16). Table 2 provides a list of words with the highest disagreement to agreement ratio, and those with the highest agreement to disagreement ratio. Thus, for example, when the word "waitress" appears in the occupation description, there are over 18 coder agreements for every disagreement. The words for which the disagreement ratio was highest are clearly more abstract and general than the words for which the agreement ratio was highest. The more abstract a term, the larger the number of legitimate interpretations.

Given that both length and the presence of certain words affected the likely coder agreement, we sought to examine the combined effects of these two factors. We found this difficult to do as some words (e.g., "assist") rarely occurred alone, while other words (e.g., "waiter") rarely occurred in combination with other words. However, an examination of a set of selected words supports the possibility of an interaction effect. For example, when the word (or part of word) "manage" appears in the description, coder agreement declines with increasing length of description. On the other hand, coder agreement is higher when the word "operate" appears with more words. Unfortunately, the QA data set was not perfectly suited to this kind of analysis. In particular, the data set did not contain enough comparable descriptions that were both long and short and involved the same easy (high agreement ratio) and difficult (low agreement ratio) words. To address this issue we carried out an experiment, which allowed us to control the characteristics of the descriptions.

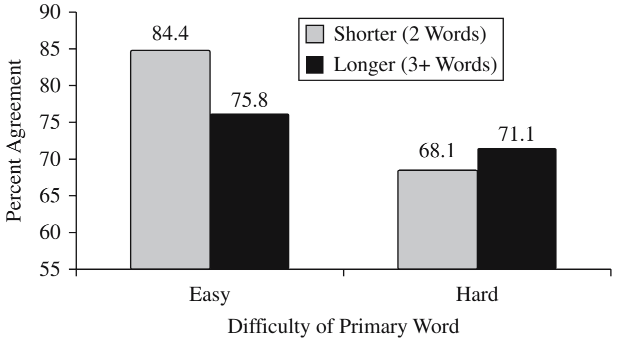


Fig. 2. Coder agreement by length of description and difficulty of primary word

4.2. Occupation Coding Experiment

The results from the experiment confirmed our earlier hypothesis that the length effect depends on the presence of certain words. When the primary word is easy, longer occupation descriptions (three or more words) decrease coder agreement, but when the primary word is difficult, longer descriptions were coded with marginally greater reliability. This is illustrated in Figure 2. When the interaction term is included in a logistic regression model, the model is statistically significant (Wald chi-square = 10.8, d.f. = 3, $p = 0.012$) and the interaction term marginally so (Wald chi-square = 2.74, d.f. = 1, $p = .098$). The secondary words' difficulty had no impact on coding reliability.

While the experimental results enabled us to elaborate on the earlier finding regarding length of occupation description, they provide little insight into *why* this pattern occurs. Our intuition was that we might gain some insight by examining coders' thinking while they make their classification decisions. Thus the last step in our investigation was to ask a small number of coders to think aloud while coding a small number of the experimental descriptions.

4.3. Coder Observations

All four coders reported following specialized coding rules that pertained to descriptions with specific characteristics or that included specific terms. The coders could not produce any of these rules in writing, nor could they provide a clear rationale for the rules or comment on their origin. These rules tended to be concerned with superficial aspects of the descriptions, rather than the concepts behind the relevant occupation or the logic of the overall classification system. Such rules are likely to increase coder agreement to the degree that they are followed by all coders on all occasions. However, there is no reason that they should improve validity, given their atheoretical character, and in fact their use may contribute to correlated coder variance (Martin et al. 1995; Sturgis 2004).

One of the rules that coders reported using could help to explain the interaction between length and difficulty that we observed in the experiment:

> When multiple occupations and multiple duties are described, select the occupation that corresponds to the duty listed first, even if it is not the first occupation.

Given the following description, this rule would dictate that the case should be coded as a driver: although "driver" is listed second in the occupation description, the corresponding duties ("delivery") are listed first.

OCC: COOK, DRIVER

DUTIES: DELIVERY, COOKING

If we assume that occupation descriptions with multiple occupations and duties are longer than descriptions with only one occupation, using such a rule could increase reliability for long descriptions, irrespective of the inherent difficulty of their component words. In contrast, for descriptions with a single (shorter) occupation, the rule does not apply; descriptions with difficult terms will therefore be less reliably coded than those with easy terms.

If such a rule is consistently applied it will improve reliability, but it may well degrade validity: in the CPS interview, respondents are first asked about their occupation, then about their duties in that occupation. This could lead to identical occupation and duty answers, which respondents might feel is uncooperative in the sense of Grice (1975). To avoid redundancy (i.e., provide different answers to the different questions), respondents may simply reverse the order of the duties relative to the occupations, irrespective of which duty best describes their occupation. Thus the order of the duties may have little substantive meaning.

Another frequently mentioned type of rule involves directly coding specific terms:

> If the word "secretary" appears in the occupation line, code to secretary and ignore all other information.

The rule would require that the following description be coded as secretary, regardless of the other information it contains:

OCC: SECRETARY, CUSTOMER SERVICE ADVISOR

DUTIES: BILLING CUSTOMERS, SCHEDULING SERVICE, ADVISING

This rule seems to be given priority over other rules, so even though the first rule for multiple occupations and duties could apply here, the direct coding rule for secretary is applied. In other cases the priority is less clear. For example, one rule was:

> If you see "assistant anything," drop the "assistant" and code to the other word.

But another rule stated

> If you see "teacher's assistant," drop the "teacher" and code "assistant."

Yet another rule applied to "assistant to . . . ," in which case the coder was to look at the duties rather than the occupation line. Hence, the rules themselves may be contradictory under certain circumstances. Furthermore, the above suggests that the rules depend on the order in which the words appear in the occupation description.

5. Conclusions and Implications

The current study of occupation coding produced three main findings. First, we found that longer occupational descriptions were less reliably coded than shorter ones. The pattern appeared to depend on the particular occupation terms involved. Second, for easy occupation terms longer descriptions were less reliably coded, but for difficult occupation terms longer descriptions were slightly more reliably coded than short descriptions.

The third main finding was that coders rely on the use of arbitrary coding rules based on superficial features of the description. It is possible coders' use of one of these rules – the rule for coding descriptions with multiple occupations and duties – could explain the interaction of length and difficulty observed in the experiment because the rule is most likely to be applicable to longer descriptions, making the difficulty of the words in those descriptions less important than in shorter descriptions. While such rules are likely to promote the reliability of coding, they are unlikely to improve validity. Although the rules are likely to be specific to a particular survey operation, the general phenomenon seems to be widespread (Hak and Bernts 1996; Campanelli et al. 1997).

We see several areas where concerted effort might directly improve the accuracy of coding in most survey operations. The first involves the training for interviewers and coders. The interviewers (who collect the occupation descriptions) can become more skilled at eliciting descriptions that are not unnecessarily long or overly specific, particularly for easy occupation terms. This should increase intercoder reliability and reduce the effect of length of description (as it was most problematic for descriptions containing easy terms). Concerning coders, if all coders are regularly exposed to a set of cases producing high disagreement (or low quality by some other measure), the group can discuss these cases and reach consensus on how to code them based on sound, theoretical reasons. This should increase agreement for descriptions that would previously have led to different codes from different coders. In addition, the informal coding rules of the sort we observed should be carefully evaluated and, if deemed to improve valid classification, should be formalized and made explicit; coders should be instructed to use them consistently. It should be possible to identify and document exceptions and develop ways to resolve conflicts among rules. Rules that are not found to improve the validity of classification should be explicitly discouraged. In fact, since the time of our coder observation in which we observed the use of many informal rules, some of these rules have been made explicit as "Job Aids" in the coder instructional materials (U.S. Census Bureau, 2014).

A second area ripe for improvement concerns the occupational coding software system used by coders. The CPS coding system in use when we observed coders suffered from numerous usability problems that could be identified and fixed with proper usability evaluation. As with any software, usability engineering can greatly improve the speed, accuracy and satisfaction of use. This is important because the way in which the software provides results from searches to coders could facilitate the application of incorrect rules. The appropriate rules used to resolve complex or ambiguous cases could be formally built into the software system.

Similarly, a more usable and flexible coding system might allow coders to assign a description that seems to belong to more than one job category to all appropriate categories, in the way that respondents are sometimes allowed to choose more than one race category to describe themselves. But this would involve a major departure from current practice about occupation data, where traditionally a job is classified just once. If jobs can be assigned to multiple categories, the number of such "composite" jobs would be vast, given that there are 501 occupations that can potentially be combined with each other in contrast to, say, race categories for which this practice would result in far fewer composite categories (see Jones and Bulock 2012).

Finally, the data collection process can be honed to improve the quality of descriptions by more fully engaging interviewers in the coding process. Campanelli et al. (1997, 450) remark that using interviewers for I&O coding may not achieve the same levels of accuracy as specialized office coders, but interviewers who are responsible for coding occupation should have a better sense of what constitutes a good occupational description and probe accordingly for more information. At the very least, the interviewers should be trained on the logic and rationale behind the coding structure so that they have a better sense of the kinds of decisions coders need to make. In addition, decision criteria can be implemented as part of the data collection software, making them available during the interview to support the coding task. For example, when the interviewer types in a term that is known to be problematic, the system would propose particular probes that should resolve the coding problems.

The current study was restricted to interviewer administration of occupation and duties questions and transcription of respondents' descriptions – as is the case in most government surveys that collect such data. But mode may matter. Self-administered questionnaires that are used to collect occupation descriptions (e.g., the American Community Survey is administered both online and by mail, as well as via telephone and personal interviews) require respondents to type or write their answers. Because writing and typing require more effort for most people than speaking, it could be the case that occupation descriptions tend to be shorter in self-administered (visual) modes. If so, the kinds of length-of-description effects we observed here might be reduced when responses are textual. This is an area – especially with the growth of online survey administration – that certainly warrants further study.

Another way in which the current study was restricted was the lack of information about individual coders, in particular their experience and competence. This might affect agreement and moderate the patterns observed here. Unfortunately we did not have access to any information about individual coders, so we could not quantify such effects. Future studies might extend the current findings by including coder information in analyses of coding quality.

Coding open-ended responses is an overlooked source of survey error. More accurate coding, rather than just more reliable coding, should be a priority. If this is achieved, then more reliable coding will follow.

6. References

Andrews, M. 2005. "Who is Being Heard? Response Bias in Open-Ended Responses in a Large Government Employee Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, Minneapolis, MN, 3760–3766.

Anker, R. 1998. *Gender and Jobs: Sex Segregation of Occupations in the World*. Geneva: International Labour Office.

Barsalou, L.W. 1985. "Ideals, Central Tendency and Frequency of Instantiation as Determinants of Graded Structure in Categories." *Journal of Experimental Psychology: Learning, Memory and Cognition* 11: 629–654.

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

Bjerk, D. 2007. "The Differing Nature of Black-White Wage Inequality Across Occupational Sectors." *The Journal of Human Resources* 42: 398–434. Doi: http://dx.doi.org/10.3368/jhr.XLII.2.398.

Bureau of Labor Statistics. 2014. "Revising the Standard Occupational Classification." Working Paper. Available at: http://www.bls.gov/soc/revising_the_standard_occupational_classification_2018.pdf (accessed September 11, 2015).

Bushnell, D. 1995. "Computer Assisted Occupation Coding." Working Paper. Available at: http://www.blaiseusers.org/1995/papers/bushne95.pdf (accessed September 11, 2015).

Campanelli, P.C., K. Thomson, N. Moon, and T. Staples. 1997. "The Quality of Occupational Coding in the United Kingdom." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, and C. Dippo, 437–457. Hoboken, NJ: Wiley-Interscience.

Cantor, D. and J. Esposito. 1992. "Evaluating Interviewer Style for Collecting Industry and Occupation Information." In Proceedings of the Section on Survey Research Methods, August, Boston, MA, 661–666.

Cawley, J.C. and G.T. Homce. 2003. "Occupational Electrical Injuries in the United States, 1992–1998, and Recommendations for Safety Research." *Journal of Safety Research* 34: 241–248. Doi: http://dx.doi.org/10.1016/S0022-4375(03)00028-8.

Couper, M.P. and F.G. Conrad. 1996. "Collecting Data to Facilitate the Classification of Occupations Using a Skill-Based Approach." Paper presented at Fourth International Social Science Methodology Conference, July, Essex, UK.

Ericsson, A. and H. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*, rev. ed. Cambridge, MA: MIT Press.

Esuli, A. and F. Sebastiani. 2010. "Machines that Learn to Code Open-Ended Survey Data." *International Journal of Market Research*, 52: 775–800. Doi: http://dx.doi.org/10.2501/S147078531020165X.

Grice, H.P. 1975. "Logic and Conversation." In *Syntax and Semantics: Volume 3, Speech Acts*, edited by P. Cole and J.L. Morgan, 41–58. New York: Academic Press.

Hak, T. and T. Bernts. 1996. "Coder Training: Theoretical Training or Practical Socialization?" *Qualitative Sociology* 19: 479–501. Doi: http://dx.doi.org/10.1007/BF02393420.

Hammond, S.K., G. Sorensen, R. Youngstrom, and J.K. Ockene. 1995. "Occupational Exposure to Environmental Tobacco Smoke." *Journal of the American Medical Association* 274: 956–960. Doi: http://dx.doi.org/10.1001/jama.1995.03530120048040.

Heywood, J.S. and P.L. O'Halloran. 2005. "Racial Earnings Differentials and Performance Pay." *The Journal of Human Resources* 40: 435–452. Doi: http://dx.doi.org/10.3368/jhr.XL.2.435.

Israel, G.D. 2010. "Effects of Answer Space Size on Responses to Open-Ended Questions in Mail Surveys." *Journal of Official Statistics* 26: 271–285.

Jones, N.A. and J. Bullock. 2012. "The Two or More Races Population: 2010." 2010 Census Briefs. Available at: https://www.census.gov/prod/cen2010/briefs/c2010br-13.pdf (accessed April 10, 2015).

Kauppinen, T., J. Toikkanen, D. Pedersen, R. Young, W. Ahrens, P. Boffetta, J. Hansen, H. Kromhout, J.M. Blasco, D. Mirabelli, V. Orden-Rivera, B. Pannett, N. Plato, A. Savela, R. Vincent, and M. Kogevinas. 2000. "Occupational Exposure to Carcinogens in the European Union." *Occupational and Environmental Medicine* 57: 10–18. Doi: http://dx.doi.org/10.1136/oem.57.1.10.

Kromhout, H., E. Symanski, and S.M. Rappaport. 1993. "A Comprehensive Evaluation of Within- and Between-Worker Components of Occupational Exposure to Chemical Agents." *The Annals of Occupational Hygiene* 37: 253–270. Doi: http://dx.doi.org/10.1093/annhyg/37.3.253.

Kunda, Z. and K.C. Oleson. 1995. "Maintaining Stereotypes in the Face of Disconfirmation: Constructing Grounds for Subtyping Deviants." *Journal of Personality and Social Psychology* 68: 565–579. Doi: http://dx.doi.org/10.1037/0022-3514.68.4.565.

Layne, L.A. 2004. "Occupational Injury Mortality Surveillance in the United States: An Examination of Census Counts from Two Different Surveillance Systems, 1992–1997." *American Journal of Industrial Medicine* 45: 1–13. Doi: http://dx.doi.org/10.1002/ajim.10308.

Lettau, M.K. 2003. "New Estimates for Wage Rate Inequality Using the Employment Cost Index." *The Journal of Human Resources* 38: 792–805. Doi: http://dx.doi.org/10.3368/jhr.XXXVIII.4.792.

Lyberg, L. and P. Dean. 1992. *Automated Coding of Survey Responses: An International Review*. R&D Reports, No. 2. Stockholm, Sweden: Statistics Sweden.

Lyberg, L. and D. Kasprzyk. 1997. "Some Aspects of Post-Survey Processing." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 353–370. New York: Wiley.

Martin, J., D. Bushnell, P. Campanelli, and R. Thomas. 1995. "A Comparison of Interviewer and Office Coding of Occupations." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, Orlando, FL, 1122–1127.

Macchia, S. and M. D'Orazio. 2001. "A System to Monitor the Quality of Automated Coding of Textual Answers to Open Questions." *Research in Official Statistics* 4: 7–21.

Moscarini, G. and K. Thomsson. 2007. "Occupational and Job Mobility in the US." *The Scandinavian Journal of Economics* 109: 807–836. Doi: http://dx.doi.org/10.1111/j.1467-9442.2007.00510.x.

Norman, D.A. 1981. "Categorization of Action Slips." *Psychological Review* 88: 1–15. Doi: http://dx.doi.org/10.1037/0033-295X.88.1.1.

Reichard, A.A. and L.L. Jackson. 2010. "Occupational Injuries among Emergency Responders." *American Journal of Industrial Medicine* 53: 1–11. Doi: http://dx.doi.org/10.1002/ajim.20772.

Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. 1976. "Basic Objects in Natural Categories." *Cognitive Psychology* 8: 382–439. Doi: http://dx.doi.org/10.1016/0010-0285(76)90013-X.

Rosenthal, N. 1992. “Evaluating the 1990 Projections of Occupational Employment.” *Monthly Labor Review* 115: 32–48.

Shniper, L. 2005. “Occupational Mobility, January 2004.” *Monthly Labor Review* 128: 30–35.

Smyth, J.D., D.A. Dillman, L.M. Christian, and M. Mcbride. 2009. “Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?” *Public Opinion Quarterly* 73: 325–337. Doi: http://dx.doi.org/10.1093/poq/nfp029.

Sturgis, P. 2004. “The Effect of Coding Error on Time Use Survey Estimates.” *Journal of Official Statistics* 20: 467–480.

Sudman, S., N. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass Publishers.

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

Tversky, A. and D. Kahneman. 1974. “Judgment Under Uncertainty: Heuristics and Biases.” *Science* 185: 1124–1131. Doi: http://dx.doi.org/10.1126/science.185.4157.1124.

Tversky, A. and D. Kahneman. 1983. “Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.” *Psychological Review* 90: 293–315. Doi: http://dx.doi.org/10.1037/0033-295X.90.4.293.

U.S. Census Bureau. 2013. Current Population Survey Interviewing Manual. Available at: http://www.census.gov/prod/techdoc/cps/CPS_Manual_June2013.pdf (accessed September 11, 2015).

U.S. Census Bureau. 2014. *Current Population Survey (CPS) and American Community Survey (ACS): Coding Instructions for 2007/2010/2012 Industry and Occupation (I&O) Coding*. Revision 2. October 1, 2014.

Westat/AIR. 1989. *Research on Industry and Occupation Questions in the Current Population Survey. Final Report to the Bureau of Labor Statistics*. Washington, DC: Westat, Inc. and American Institutes for Research.

Received February 2014
Revised September 2015
Accepted October 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 93–112, http://dx.doi.org/10.1515/JOS-2016-0004

Census Model Transition: Contributions to its Implementation in Portugal

Carlos A. Dias[1], *Anders Wallgren*[2], *Britt Wallgren*[3], *and Pedro S. Coelho*[4]

Given the high cost and complexity of traditional censuses, some countries have started to change the census process. Following this trend, Portugal is also evaluating a new census model as an alternative to an exhaustive collection of all statistical units. The main motivations for the implementation of this census model transition in Portugal are related to the decrease in statistical burden on citizens, improvements in the frequency of outputs, and the reduction of collection costs associated with census operations. This article seeks to systematise and critically review all alternatives to the traditional census methodologies, presenting their advantages and disadvantages and the countries that use them. As a result of the comparison, we conclude that the methods that best meet these objectives are those that use administrative data, either in whole or in part. We also present and discuss the results of an inventory and evaluation of administrative registers in Portugal with the potential to produce statistical census information.

Key words: Administrative registers; register-based census; traditional census.

1. Introduction

Population and housing censuses are statistical operations performed across the world to collect all data on the statistical units – living quarters (dwellings), households, and persons – within a national universe. Traditionally, census operations are decennial and require significant human, financial, and material resources. In addition to these high costs, a considerable effort is required from citizens, who are "forced" to respond to questions whose answers, in many cases, may already exist in several databases within the Public Administration (Scheuren 1999). Given these constraints, some countries have started to change the census process. This entails collecting data not from the traditional model involving an exhaustive survey of all statistical units, but from administrative sources (Redfern 1986).

This article aims to present a systematic critical review of alternative methodologies to the traditional census, showing their advantages, disadvantages and the countries in which they are used. The methods using administrative data are highlighted and an analysis of

[1] Nova IMS-Information Management School and Statistics Portugal, Av. António José de Almeida, 5, 1000-043 Lisboa, Portugal. Email: carlos.dias@ine.pt
[2,3] BA Statistiksystem AB, Norrhagev.30 SE-71993 Vintrosa, Sweden. Email: ba.statistik@telia.com
[4] Nova IMS-Information Management School, Campus de Campolide, 1070-312 Lisboa, Portugal. Email: psc@novaims.unl.pt

Acknowledgments: The first author would like to thank is Fernando Casimiro, head of the Census Unit at Statistics Portugal, for his helpful suggestions and comments. The views expressed in this article are those of the authors and do not necessarily reflects the policies of Statistics Portugal.

their situation in Portugal is included. The objective is to lay the foundations for the use of administrative data, in whole or in part, in the country's 2021 census round.

This article is organised in seven sections. Section 2 briefly presents the legal framework at national and international level. Section 3 identifies the methods used to obtain census information and explores a critical review of alternative methodologies to the traditional census. Section 4 discusses the census model transition in Portugal. Section 5 analyses the administrative registers in Portugal of potential use for statistical purposes in the census context. Section 6 discusses possible models for the 2021 Census in Portugal. The main conclusions of the article are presented in Section 7.

2. Legal Framework

In order to ensure the harmonisation and comparability of results, the UNECE (United Nations Economic Commission for Europe) recommendations set out the basic rules to be followed in population and housing censuses. They also establish the concepts and definitions associated with the statistical units and variables to be observed (UNECE 2006). Based on these recommendations, for the 2010 census round (covering the period 2005–2014), the EU (European Union) adopted four regulations proposed by Eurostat (Statistical Office of the European Union) after discussion with the representatives of all member states. The regulations introduce a mandatory set of rules on the content to be observed, the geographical breakdown for each variable and the quality indicators that each member state should provide to Eurostat (Eurostat 2011). The existence of this community legislation, as an instrument for regulating the EU censuses, guarantees the availability and harmonisation of census information.

In Portugal, census operations are also supported by specific national legislation. A specific law established the organisational arrangements and executive for the 2011 Census, the last traditional census operation held in Portugal. A feature of this legislation is the explicit reference to the implementation of the census transition process, supported by data from administrative sources. To this end, it includes the possibility for Statistics Portugal, which was responsible for implementing the 2011 Census, to create databases for individualised registers of living quarters/buildings, housing units/dwellings, households, and persons. This legal framework at national level also underlies the Portuguese Statistics Act, which regulates the National Statistical System (SEN) and establishes, for the first time, the principle of the use of administrative data for official statistical purposes.

3. Alternative Methodologies

According to the international recommendations for obtaining census statistics, various approaches can coexist in data collection, covering a wide spectrum from the exhaustive collection of all statistical units (traditional method) to models based solely on administrative information (register-based censuses). The mix of the two methods, supplemented in some cases by sample surveys, allows several combinations.

Numerous studies have presented different method classifications (Valente 2010b; UNECE 2013). In this article, we propose four main groups of methods that build on the classification in Valente (2010b): traditional census, register-based census, combined methods, and rolling census (Figure 1), which are detailed in the following subsections.

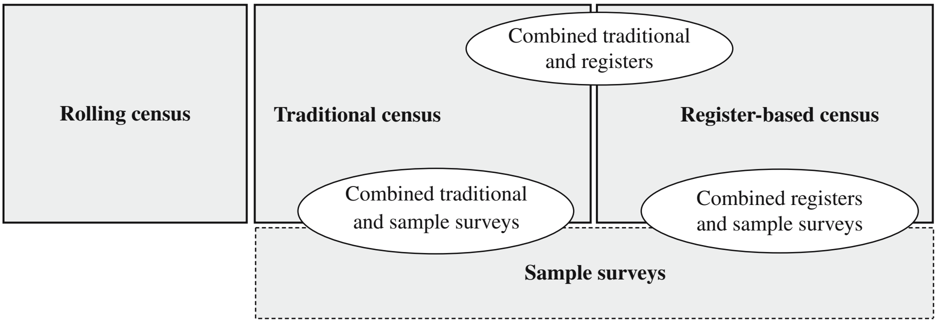

Rolling census
Traditional census
Register-based census
Combined traditional and registers
Combined traditional and sample surveys
Combined registers and sample surveys
Sample surveys

Fig. 1. Methods of obtaining census information for statistical purposes

3.1. Traditional Census

The traditional census approach collects basic characteristics from all individuals and housing units (full field enumeration) at a specific point in time. In most countries, including Portugal, this is the most common approach to census taking. In the 2011 Census, the census could be taken through a web questionnaire in 13 UNECE countries (Valente 2010b). It is noteworthy that in Portugal approximately 50.5% of the population responded via the Internet.

138 countries responded to the survey conducted worldwide by the UNSD (United Nations Statistics Division) and the UNECE in 2009 concerning the 2010 census round, and of these 83% planned to use the traditional method (UNSD 2010). Of all the UNECE member countries, about 56% of the 50 responses indicated the traditional method as the method to be adopted, while in the 27 EU member states, only 41% indicated this method (Valente 2010a). These results indicate that the most highly industrialised and developed countries are more likely to abandon the traditional method and adopt new alternative census designs.

One of the main disadvantages of the traditional method is the high cost and complexity of census operations regarding the short-term recruitment of a large number of enumerators to carry out the field work. These costs represent about 50–60% of the operation's total budget (Valente 2010b). Another problem associated with this method is presented by the increasing difficulty in conducting population surveys based on field data collection. For security reasons, many citizens, especially the elderly and those living alone, refuse to open the doors of their home to enumerators (Valente 2011). On the other hand, the rate of change in modern societies increases the demand for statistical information and the need for more frequent updates than traditional censuses allow.

3.1.1. Traditional Method with Long and Short Forms

The traditional census approach may also include the use of long and short forms, which can ease the burden on respondents and reduce the cost of census operations. The short form, with wider coverage (majority or all of the population), is intended to collect basic information on the characteristics of housing and/or population (e.g., place of usual residence, sex, age and number of household members). The long form is more detailed and only answered by a sample of the population, usually between 10 to 20% of the total potential respondents. According to the UNSD, 14 countries, including Brazil, Canada, Mexico, Russia, and the United States (up to the 2000 Census) use or have used a mixture of short and long forms (UNSD 2010).

3.1.2. Traditional Method with Sampling Annual Updates

This method combines the traditional model, carried out at intervals of five or ten years, with sampling-based annual updates. In the census year, the entire population of the country is enumerated using a short form that collects only basic socioeconomic and demographic characteristics. In the intercensus years, annual sample surveys are conducted with more detailed questionnaires (long questionnaire).

After the 2000 Census, the United States started to use this method, which included the exhaustive census, using a short questionnaire directed at the whole population and repeated every ten years. In the intercensus years, an annual survey is carried out with a larger number of variables – ACS (American Community Survey) – covering approximately two percent of the total population, which in 2010 was about 320 million (Herman 2008; Woods 2009). The ACS was fully phased in by 2005.

When compared with the decennial census, this method has the advantage of being able to provide results with greater frequency and timeliness. In exhaustive census years, it also reduces the complexity and burden involved in the use of the long form. Furthermore, in the intercensus period, it allows for the methods and techniques of sample surveys to be developed and readjusted. However, this method has certain disadvantages: the high financial costs of carrying out the surveys annually and the technical complexity of the associated procedures, especially the construction of the estimators; moreover, the data on detailed characteristics are limited since they only come from a sample survey (the ACS).

3.2. Register-Based Census

The register-based census method does not use field operations and forms to collect data. The census statistical information is produced solely and exclusively based on administrative data, which are updated regularly according to input information from administrative acts carried out by the population on a daily basis.

The Nordic countries were pioneers of this method. Denmark was the first country in the world to move from the traditional census to a register-based census based entirely on administrative registers. The long-term work and strategy behind the first register-based census in 1981 is described in a book by Statistics Denmark (1995). In Finland, the use of administrative records for statistical purposes began as early as 1970 and has increased since that date (Statistics Finland 2004). Since 1990, the census has been based entirely on information in the registers without using a single form (Myrskylä 1991). In Norway and Sweden, as well as in Austria, exhaustive census operations have also been abandoned and the 2011 Census was fully supported by administrative data (Tönder 2008; Andersen and Utne 2011; Berka et al. 2010).

It is estimated that in Finland, the costs of conducting the census through the register-based census method were less than a tenth of the cost of using a classical method with postal questionnaires but without enumerators (Statistics Finland 2004). However, it should be noted that the costs required for establishing and maintaining a register-based statistical system can be very significant. Other benefits associated with this method are that the information processing is faster than the traditional method, which also implies greater speed in the delivery of results. The main disadvantages in using this method result from restrictions on access to administrative data and their limitations in terms of content and quality. The variables used are defined by the administrative needs and rules of the organisations that produce them and not from a statistical perspective. Administrative data are often incomplete, inconsistent, outdated, or limited in their coverage. In many areas, some of the mandatory variables, included in international recommendations, may be difficult to obtain or impossible to find. Moreover, administrative data are focused on individuals and generally provide limited information on families, which limits social analysis (Dugmore et al. 2011; Zhang 2011). Finally, the concepts and classifications associated with the variables may not correspond to the statistical concepts that must be observed.

For a system of administrative registers to work effectively, it is necessary to ensure the links between the different records, which is generally possible with a unique identification key. This should correspond to an identifier which is not subject to mutations over time and that unambiguously identifies only one statistical unit. In Denmark a crucial component was the public administration's introduction of a fixed personal identification number for each individual, which replaced the different identifications previously used (Borchsenius 2000). Equally essential was the coding of addresses, which are considered a key link in the whole system. These are assigned a unique number (address code), thus allowing an interconnection between, for example, the Central Population Register and the Buildings/Dwellings Register. In countries without address codes, other methodologies of matching files are under investigation (Maldonado et al. 2010; Winkler 2011; Conti et al. 2012; Zhang 2012).

3.3. *Combined Methods*

Some countries obtain the census information through a combination of methods, designated as combined or mixed methods.

3.3.1. Traditional Method Using Administrative Registers

Some countries use administrative information to improve the accuracy of enumerations and the quality of data. They use address lists to support field operations and may send the questionnaires to respondents via mail. Part of the questionnaire may already be completed with data obtained from administrative sources (e.g., housing address, occupant names, sex, date of birth, etc.). The respondents or enumerators (through direct interview) only correct or update the information and complete the remaining questions. The average time of interview or completion is substantially reduced, which implies a positive impact on costs and improves the quality of the data. In addition, coverage can be evaluated by comparing the population register and fieldwork results.

When compared to the register-based census model, this method is more expensive, complex in its implementation and increases the burden on respondents (Redfern 1989). Some countries used this model in 2011: the Czech Republic, Estonia, Latvia, Lithuania, and Italy (Valente 2010b). Spain also considered using it (Ballano 2008), though it finally opted for a combined method of administrative records, exhaustive collection of information on buildings, and sample surveys of dwellings and population.

3.3.2. Administrative Registers and Sample Surveys

As the administrative registers do not contain all the information required, some countries complement registers with sample surveys. The registers are used to ensure that the entire population is counted and the survey results allow the missing individual characteristics to be obtained. The surveys can be designed specifically for the census (ad hoc surveys) or may already exist.

A mixed method, using administrative registers combined with sample surveys already in use, was implemented in the Netherlands for the first time in the 2001 Census, and was also adopted in 2011. This model, known as the Virtual Census (Nordholt 2005), does not require specific field operations but implies a complex estimation process for the lower levels of breakdown (Houbiers 2004), links between the records (Linder 2004; Nordholt and Linder 2007) and strict quality control (Daas et al. 2009; Nordholt et al. 2011). Slovenia also adopted this model in 2011 (Dolenc 2010). Germany complements this model with additional specific surveys (Eppmann et al. 2006; Szenzenstein 2005). Even using existing surveys, the model can also be combined with ad hoc surveys to evaluate the accuracy and degree of record completion or to add new variables (such as in long forms). Israel adopted this approach in the 2008 Census, thereby improving the accuracy of population registers and adjusting their counts (Valente 2010b).

3.4. Rolling Census

France is the only country that uses the rolling census method, first proposed by Kish (1986; 1990). Implemented in 2004, it is based on annual surveys, which each year cover about 14% of the total population in parts of the country during a five-year cycle. The nearly 37,000 communities in France are classified into two groups: small and medium-sized communities (with fewer than 10,000 residents) and large communities (with 10,000 or more residents). Small and medium-sized communities, containing about half of the country's total population, are divided into five groups. Every year during the cycle, in rotation, each group is subject to an exhaustive census of all dwellings and people. In large communities, during the cycle, a sample survey, covering about eight percent of the dwellings, is held annually. At the end of five consecutive years the whole population of small and medium-sized communities and approximately 40% of the population of large communities has been surveyed. Overall, about 70% of the French population is covered during the entire lifecycle (Durr and Dumais 2002).

The advantages of this method are the possibility of distributing the efforts and costs over five years and improvements in the frequency of results – annual results in contrast to classical methods. The major disadvantage is the mobility of the census moment, which implies that data are not collected simultaneously for the whole population. Even if the data collected are adjusted to the same period, this poses certain difficulties in comparing areas surveyed at different times. The respondents' mobility over the five years also has implications in the model – it can cause gaps or duplications in the population. It also has the disadvantage of involving a highly complex methodological approach, especially with the use of sampling techniques and modelling.

4. Census Transition in Portugal

In recent years, Portugal has been promoting policies that will modernise public administration services. In 2006, one of the most visible consequences of this strategy was the creation of SIMPLEX, a national governmental programme for administrative and legislative simplification that aims to improve and facilitate the interaction of citizens and businesses with public administration. One example of such advancements is the IES program – Simplified Business Information. This enables enterprises to reply to the Public Administration only once, through an electronic form, replacing several surveys that contained the same questions and were collected by different entities (Ministry of Finance, Bank of Portugal, Statistics Portugal, etc.). With a view to continuing the modernisation of the data collection process, the Action Programme for the 2011 Census provided the assessment of administrative registers for statistical purposes and a methodological review (INE 2010).

As mentioned above, a fundamental aspect resulting from other countries' experience of the use of administrative data is the existence of specific legislation that allows national statistical institutes access to these data (Wallgren and Wallgren 2007). For the first time in Portugal, the National Statistical Act allows Statistics Portugal to access individual administrative data collected by public sector entities. Following the trend of other UNECE countries, Portugal is thus also able to evaluate a new census model based on administrative data. The need for census model transition is underpinned not only by the high financial resources allocated to traditional census operations, but also the enormous effort required of citizens every ten years. Accordingly, the main motives for the transition are focused on contributions to society: to decrease the burden on citizens, to allow for a greater frequency of census data (annual) and to reduce the high costs associated with census operations. According to the description of census methods presented in Section 3, the methods that best fit these goals are those based on administrative data: the register-based census and combined methods. Their adoption reflects a change of paradigm in census operations in Portugal because it involves (re)thinking the approach and methodological design associated with the production of statistics based on administrative sources. Furthermore, it requires coordination between the different entities that produce and manage the administrative data (Statistics Portugal 2010). Thus the problem is to define a new methodology, based wholly or partly on administrative data, so as to replace the traditional population and housing censuses in Portugal. However, in order to achieve this objective, we must first find answers to the following questions:

- What administrative registers are available and what information do they contain?
- Does the information that exists in administrative data meet the requirements of census users, international recommendations and EU regulations?
- Do the available variables correspond to the fundamental questions of the census in terms of coverage, content and quality?
- What gaps exist in terms of census variables and what methodologies should be implemented to obtain the desired information?

In order to answer these questions, it will be necessary to evaluate the existing administrative registers in Portugal and their potential to produce census statistics.

5. Administrative Registers in Portugal

One of the main dimensions to consider when evaluating a transition model is whether there are administrative files and registers with individual data (microdata) and unique identification that will be of interest to the census. In countries that have implemented systems of administrative records for statistical purposes, the combination of different sources was also the key factor in the process. To achieve this combination, the quality of the various sources was assessed by comparing and validating the information (UNECE 2007; Wallgren and Wallgren 2007). Following these principles, the research sought to identify administrative sources covering the statistical units used in the census: housing (buildings and dwellings) and population (households and persons). The first step therefore consisted of identifying and assessing sources with statistical units of housing. The second phase of activities focused on population data. Table 1 shows the main administrative registers identified as having potential for the mandatory census variables required by international recommendations, as well as the entities that manage them.

The records in the potential files of interest must be evaluated in terms of coverage, content, quality and identifiers. This task involves the analysis of:

- Individual records, the level of harmonisation, standardisation and consistency of information collected by different entities,
- Updating and management systems,
- Metadata associated with the data.

It is important to note that the information collected by administrative entities does not necessarily correspond to the statistical concepts to be observed according to international recommendations. In order to be used as statistical data, administrative registers undergo several transformations: coding of variables and creation of derived variables and validations, among others (Wallgren and Wallgren 2007).

Under the current legislation, Statistics Portugal has access to some administrative registers with potential for characterising census variables. As far as housing units are concerned, Statistics Portugal has access to the real estate register (municipal property tax), income register (personal income tax) and energy register.

The access to administrative registers of individuals has posed some difficulties. Based on different interpretations of the current Portuguese statistical act, register managers have been reluctant to share their registers. They only allow access when the Portuguese Data Protection Authority gives its consent. Statistics Portugal has assessed, among others, the civil register, the social security register, the employment register and the foreigner register.

The following sections present the preliminary results of the analysis of available administrative records for housing and population statistical units.

5.1. Housing

With regard to housing information, the real estate register is the core register among the identified relevant administrative registers. It is the most extensive register available in

Table 1. Relevant administrative registers, register holders and content

Administrative registers	Register holder	Content
Real estate register -Municipal property tax	Tax and Customs Authority of the Ministry of Finance	Buildings and dwellings characteristics
Energy register	Energy register – Portugal Energy	Household addresses of electric energy consumption
Civil register	Registers and Notaries Institute of the Ministry of Justice	Some census variables related to population
Social security register	Information Institute of the Ministry of Solidarity and Social Security	Potential interest of some socioeconomic variables
Employment register	Strategy and Planning Office of the Ministry of Solidarity and Social Security	Information about workers under individual contract of employment
Income register - Personal income tax	Tax and Customs Authority of the Ministry of Finance	Personal tax addresses; household structure
Foreigner register	Foreigner register – Foreigners and Borders Office, Ministry of Internal Administration	Characterisation of the foreign population
Unemployment register	Unemployment register – Employment and Professional Training Institute, Ministry of Economy and Employment	Data on registered unemployed
Public administration register	Public administration register – Directorate General of Public Administration of the Ministry of Finance	Data on public administration staff
Pensioner register	Pensioner register – Pensioners Office of the Ministry of Finance	Data on pensioners
University register	Directorate General of Statistics and Science	Higher education administrative data
School register	Directorate General of Statistics and Science	Education administrative data (not higher education)

Table 2. Available EU mandatory census variables related to housing

Variables available in the real estate register	Variables not available in the real estate register
Type of living quarters	Housing arrangements
Location of living quarters	Occupancy status
Useful floor space and/or number of rooms	Type of ownership
Water supply system	Number of occupants
Toilet facilities	Bathing facilities
Type of heating	Density standard
Type of building/number of dwellings	
Period of construction	

regard to coverage, content and identifiers. Real estate register data are produced by administrative acts covering municipal property tax. All properties (buildings or parts) located in Portugal that pay these taxes are included. As shown in Table 2, the real estate register covers around 57% (eight out of 14) of the mandatory variables of the EU regulations for the 2010 Census Round (Eurostat 2011; UNECE 2006) concerning building and dwelling characteristics. It will not be possible to ascertain the housing arrangements, occupancy status, type of ownership, number of occupants, bathing facilities and density standard.

Moreover, there are significant differences in terms of concepts and categories associated with the real estate register variables and the census variables required. The concepts of *housing property* and *fraction*, used in tax administration, are different from the census statistical units *building* and *dwelling*. In addition to these problems, we identified limitations in terms of harmonisation of fields and low rates of completion of some variables. To complete the missing information, other potentially useful files must be found; alternatively, the possibility of including new fields or making changes to existing forms associated with the real estate register should be evaluated in collaboration with the register managers.

The analysed data file from 2010 (referenced to December 31) only covers administrative acts carried out between 2003 and 2010. The housing properties of this flow of data only represent around 36% of total dwellings obtained in the provisional results of the 2011 Census. Statistics Portugal also needs to request the real estate register keepers'/managers' permission to access the global stock of housing properties. In terms of housing topics, the energy register (2008) and the income register (2008) were also analysed. These registers do not provide census information, thus their potential is limited to the use of additional information to update the housing addresses.

5.2. Population

As for population variables, about 92% (22 out of 24) of the mandatory census variables of EU regulations (Eurostat 2011; UNECE 2006) are represented in some of the existing registers (Table 3).

Some of these variables can only be obtained through the combination of two or more registers. The data present important limitations in terms of content (suitable concepts) and coverage. For example, with data from the income register and the civil register it is

Table 3. Available EU mandatory census variables related to population

Variables	Civil register	Social security register	Employment register	Income register	Foreigner register
Place of usual residence	x	x		x	x
Location of place of work or school			x		
Sex	x	x	x	x	x
Birth date (age)	x	x		x	x
Legal marital status	x	x		x	
Current activity status		x			
Occupation			x		x
Industry, branch of economic activity		x		x	
Status in employment		x	x		x
Educational attainment			x		
Country/place of birth	x	x			x
Country of citizenship	x	x	x		x
Ever resided abroad, year of arrival					
Place of residence before census year	x				
Relation between household members				x	
Tenure status of households					
Total population	x				x
Locality	x				
Household status				x	
Family status				x	
Type of family nucleus				x	
Size of family nucleus				x	
Type of private household				x	
Size of private household				x	

possible to construct some private households and family nuclei, but these registers do not cover the people exempt from income tax, such as persons with incomes lower than a defined value. For tax purposes, a young adult over 18 years working and living at his or her parents' house (sharing food and possibly other essentials for living) is not included on the same tax form (defining an income register household) as his parents. Instead, he or she is represented with a separate tax form. In census concepts, a young adult and his parents constitute a single household. At the moment, Statistics Portugal is unable to access the complete file of the income register. Some attempts have been made to gain access to these administrative data but so far without success.

The civil register file should be the core register for the population statistical units. However, some population groups are not represented in the civil register file: the foreign resident population (which has no Portuguese civil identification) with the exception of Brazilian citizens with equal status (resulting from the Treaty of Porto Seguro agreed between Portugal and Brazil), and children who were born before 2007 and do not have the

Citizen Card (CC). The “Born Citizen” project, introduced in 2007, means that when children are born and registered they receive a civil identification number even if this is not requested (the CC is only compulsory from six years of age). Those numbers will stay the same throughout their entire life. The only exhaustive coverage of all people legally residing in Portugal is the combination of the civil register and foreigner register files. However, none of these administrative registers covers the illegal population. In contrast, the traditional census gives us a global “picture” and all individuals, regardless of their legal status, should be enumerated. On the other hand, many people who are listed in the civil register with a Portuguese residence effectively reside in another country and are not enumerated in a traditional census. In the 2011 Census, an individual is considered resident when he “lives in his usual place of residence in the twelve months preceding the time of the census with the intention to stay for a minimum period of one year” (UNECE 2006).

The 2009 and 2010 civil register (referring to December 31) data files were analysed. A set of strengths and points that require better assessment for their potential use were identified for the 2010 data:

- The file presents a low percentage of null records (missing data) and a primary numeric key for all records without duplication;
- The address fields of the individuals who already have the CC are standardised and are of good quality, which may allow them to be matched with other sources;
- The administrative division classification maintains the existing code on the date of occurrence. Thus the codes for geographic places of birth and residence are not adjusted to the administrative division changes that have occurred over the years, implying, for example, the existence of the same codes for different administrative areas;
- It is clear that bilateral cooperation between Statistics Portugal and keepers/managers is important in order to improve the quality of civil register data through standardisation and joint verification of discrepancies.

The foreign population residing in Portugal was enumerated in Census 2011 but does not appear in the civil register files. The total resident population in Portugal obtained in the provisional results – the reference date was the census moment (21 March 2011) – was 10,561,614. The civil register (on 31 December 2010) presents 7.1% more resident persons than the 2011 Census. The two sets of data were not collected on the same date. For a correct comparison it would be necessary to make adjustments for the same reference date. However, since the gap is less than three months and the Census 2011 results are provisional, we do not make the adjustments here.

The protocol signed between Statistics Portugal and Social Security permits annual access to the social security register data. The stock of 2008 to 2010 (referring to December 31) social security register data files was analysed. The data present quality for use: they contain an exhaustive primary key (Social Security Identification Number) for all registers; the variables are standardised and present a low percentage of missing data; the percentage of inconsistent data is small; and the classifications are those used by Statistics Portugal. According to the keeper/manager of these administrative registers, the gaps and incoherencies (e.g., geographic codes) in some variables are the result of a process which occurred in the late 1990s. The Ministry of Solidarity and Social Security gathered all regional registers into one centralised database without a verification or confirmation process. Currently, any update or correction of data concerning beneficiaries is automatically registered in the database and subject to an intense validation process, contributing to increased quality of data. The numeric field NIF (corresponding to the Tax Identification Number) presents a fill rate of around 96%. Thus this field could be used as a key liaison with other administrative files.

The employment register is obtained from the administrative data submitted annually by all employers with workers employed under individual contract of employment. The online application used to capture information establishes a set of validations that guarantee overall quality and consistency of this file. The stock of 2009 (referring to 31 October) employment register data files was analysed. The data present quality for use: they contain an exhaustive primary key (NISS – Social Security Identification Number) for all registers; the variables are standardised and present a low percentage of missing data; and the classifications are those used by Statistics Portugal. Later, it was found that the employment register keeper/manager has other available mandatory variables that have not yet been provided: date of birth and branch of economic activity.

Finally, the foreigner register can ensure the coverage and content of a very specific portion of the population (foreigners legally residing in Portugal), completing the information in the civil register files. The stock of 2009 (referring to 31 December) of the foreigner register was analysed. The data present some problems for use for statistical purposes: the identification key corresponds to a code number assigned by the foreigner register, which prevents matching with other files; the geographical breakdown only exists at municipal level; and some fields are not validated, notably in the address fields. Later, it was found that the foreigner register keeper/manager had not provided other available mandatory variables: country of last residence, Social Security Identification Number and Tax Identification Number, which could be the key liaison with other administrative files such as the social security register and employment register.

5.3. *Integration*

The development of an information system to store the individual registers of housing and population must be part of the transition census process. This can be based on existing models in other countries (UNECE 2007; Wallgren and Wallgren 2007), adapted to the Portuguese reality. For statistical purposes, this system shall include the national files of dwellings and persons, which may be based on the 2011 Census microdata and subsequent updates with housing and population administrative data.

As already mentioned above, an important part of administrative information is dispersed across multiple files managed autonomously by different entities. In addition to storage issues, the various registers within the system must be linked if the administrative registers are to be transformed into statistical records.

The geographical key location of buildings and dwelling, to be defined, could be address type (full address information and postcode), coordinate ID type (X, Y) or a combination of both. This key is fundamental in linking housing registers to population registers. The real estate register uses geographical coordinates and address.

The structure of the National Population Register for statistical purposes based on administrative data requires a unique key. This key, which we shall call the administrative key, is of great importance since it will ensure that all the administrative information on the population can be successfully integrated. The key would ideally be composed of only a single field, but due to the complexity of linking files (already experienced in the analysis), we cannot rule out the need to develop a composite key. There are files that, due to the absence of common fields, can only be related through a key from a third file. Table 4 presents the possible fields that might allow the links between files. For example, the numeric field NIF (Tax Identification Number) could be used as a key liaison between the income register, the social security register and the real estate register. In the same way, the numeric field NISS (Social Security Identification Number) could be used as a key liaison between the civil register and the social security register. In addition to the possible connections through the numeric codes NIC (Civil Identification Number), NIF and NISS, other possibilities could be tested, since the variables are standardised. For example, "Address" connects statistical units of housing to units in the population associated with aggregated "Name and Date of Birth" or "Address".

Table 4. Identities in registers that can be used as matching keys

Administrative registers	Name	Address	Geo code	NIC	NISS	NIF
Variables related to housing						
Real estate register		Address	Geo code			NIF
Energy register		Address	Geo code			
Variables related to population						
Civil register	Name	Address		NIC		
Social security register		Address		NIC	NISS	NIF
Employment register					NISS	
Income register		Address				NIF
Foreigner register	Name	Address				

NIC Civil Identification Number.
NISS Social Security Identification Number.
NIF Tax Identification Number.

6. Possible Models for the 2021 Census in Portugal

To prepare for the 2021 Census, a model for use could be established by comparing the results of the 2011 Census with existing administrative data. To avoid an exhaustive comparison of all administrative data and 2011 Census data, it should be possible to implement a test structure: choosing a number of areas in the country (e.g., a sample of municipalities) to compare data collected from the 2011 Census with the corresponding administrative data at the micro level of each statistical unit. The comparison with existing administrative records could provide answers to the following question: if the 2011 Census had not been carried out using the traditional model, could we have obtained consistent and relevant data on population and housing through administrative sources?

As identified in Section 4, the methods that best fit the objectives defined for the Portuguese census transition are those based on administrative data: *register-based* *census*, or *administrative registers and sample surveys*, or *traditional method using administrative registers*. These different approaches to census taking involve trade-offs between: overall quality of census information; the cost, complexity, and frequency of census data; and burden on citizens required for change. The register-based census is the first priority model to develop for the post-2011 Census in Portugal because of the low cost, the high frequency of data, and the fact that it is no burden on individuals. However, there may be difficulties in moving directly from a traditional to a complete register-based census as a result of the problems identified in the preliminary analysis of existing registers (performed in Section 5) and the lack of experience in the use of administrative registers for statistical census purposes. Only Austria has passed directly from a traditional census in one round (2000) to a register-based census in the next round (2010) (Valente 2011; Ralphs and Tutton 2011). This transition has generally taken several decades, according to the experience of the Nordic countries.

The second priority could be the combined method of administrative registers and sample surveys because of the reductions in cost and burden on individuals, when compared to the traditional approach. In order to implement it and complement the missing information in administrative registers, it will be necessary to evaluate all existing surveys in the Portuguese statistical system. The integration of administrative data and sample survey data requires a complex process of estimation and calibration, especially for areas with lower levels of disaggregation. Thus it will be necessary to evaluate and adapt existing models (Zanutto and Zaslavsky 2002; Houbiers 2004; Mulry et al. 2006).

If the second priority cannot be implemented, an alternative model to be developed for the 2021 Census could be the traditional method using administrative registers. With this method, all individuals are enumerated but, as explained in Subsection 3.3.1., the use of register data increases the efficiency in field operations: mailout of the questionnaires to all households in the list of households and addresses (obtained from administrative registers) and multichannel collection of responses (web, mail back, municipal office of collection). This approach may help to improve the coverage and quality of the registers and, as a result, it is often selected as the first step from a traditional census towards a register-based one (Valente 2011).

In order to implement a census based on administrative data, some of the problems identified in Section 5 need to be solved: incomplete coverage of the housing stock, excess of population in the civil register, lack of coverage of the illegal and foreign population, variables with small coverage of the population, incompatible identifiers and failure to access the complete income-tax records. Some proposals are presented to solve these problems. One fundamental element of a system for integrating administrative registers is the availability of a definitive National Address Register, providing a list of all housing addresses. To avoid the difficulties in accessing the global stock of housing properties, the housing data from Census 2011 could be the basis for the National Address Register. Additionally, it will be possible to check the coverage in the period 2003–2010 by comparing census housing data with the real estate register. If the results present a good coverage, the Address Register will be updated by the real estate register.

The problem of the excess of population in the civil register could be solved in the same way. The population data from Census 2011 could be the basis for the National Population

Register. If we compare the individual Census population data with the civil register, it will be easier to understand the differences between the two files in recent years. The results of this analysis may help to improve the quality of civil register data and evaluate the use of the civil register to update the National Population Register. However, the civil register update has more weaknesses than the real estate register updates. It will always be a complex challenge to detect people who are legally resident in Portugal and have a Citizen Card (included in the civil register) but who actually reside abroad all year and just spend vacations in their dwellings in Portugal.

The lack of coverage of the illegal and foreign population and variables with small coverage of the population could be overcome by implementing ad hoc sample surveys specifically designed for this purpose. The problem of incompatible identifiers between administrative registers could be minimised using the address as a key connection, since this field is present in almost all registers. Despite the difficulties of using this field for linking registers, new matching techniques developed in recent years have produced very robust results (Maldonado et al. 2010). In addition to solving the failure to access the complete income-tax records, it will be necessary to revise the National Statistical Act. The new legislation should be strong, clear and unambiguous, giving Statistics Portugal unrestricted access, for statistical purposes, to administrative data on unit level with identification data and the possibility to link them with other administrative registers. In order to accomplish the census information system for the transition, it will be necessary to gain access to more administrative sources, particularly in the areas not yet covered: education and unemployment.

7. Conclusions

Decennial census operations are important and require large human, financial and material resources. Given these constraints and bearing in mind that statistical information is essential, the implementation of the census transition in Portugal is focused on three goals: to decrease the burden on citizens, to allow for a greater frequency of census data (annually if possible) and to reduce the high costs associated with census operations. This article presents a systematic critical review of alternative methodologies to traditional censuses, identifying their advantages and disadvantages as well as the countries that use them. Comparing several methods, it appears that those that best fit the objectives defined for the Portuguese case are the methods that rely on administrative data. However, it is also clear that the present legal framework and the nature and quality of available administrative registers still require changes or improvements in order to enable such methodologies.

Under the current Portuguese Statistics Act, Statistics Portugal has access to some administrative registers with individual data that have potential for obtaining census variables related to housing and population. With regard to housing topics, a core register has been identified – the real estate register file (municipal property tax). The real estate register has great potential for use, although there are differences in concepts, limitations in terms of harmonisation of fields and low rates of completion of some variables. The real estate register contains information flows and does not include the stock of buildings/properties. The data analysed between 2003 and 2010 represents around 36% of total dwellings obtained in the provisional results of the 2011 Census.

As for population variables, around 92% of the mandatory census variables are represented in some of the existing registers. However, the administrative registers identified present important limitations in terms of content (suitable concepts) and coverage. The civil register should be the core register for the population statistical units. Although they belong to different universes, we compared the total resident population in Portugal, obtained in the provisional results of the 2011 Census, with the civil register file (on 31 December 2010). The civil register file presents 7.1% more of the resident population than the 2011 Census. The social security and employment administrative registers (referring to 2010) were analysed and showed good quality for use in producing census statistical information. On the other hand, the analysis of the results of the foreigner register (referring to 2009) presented significant limitations in their use for statistical purposes.

The different approaches to census taking based on administrative data involve trade-offs between the overall quality of census information, the cost, complexity, and frequency of census data, and the burden on citizens required for change. The register-based census is probably the first priority model to develop for the post-2011 Census in Portugal. However, there may be difficulties in moving directly from a traditional to a complete register-based census as a result of the problems identified in the preliminary analysis of existing registers. The second priority could be the combined method of administrative registers and sample surveys. Nevertheless, the implementation of this model implies a complex process to integrate administrative data and sample survey data. An alternative model could be the traditional method using administrative registers. This approach enables efficiencies in field operations and may help to improve the coverage and quality of the registers. Therefore, this could be considered a first step in the right direction, contributing towards a future register-based census.

It was also possible to identify some key points in evaluating administrative registers that constitute challenges for the continuity of the work and for the strategy to be defined: incomplete coverage of the housing stock, excess of population on the civil register, lack of coverage of the illegal and foreign population, variables with small coverage of the population, incompatible identifiers and failure to access the complete income-tax records. In the article, some proposals have been presented that may contribute towards solving or minimising these problems.

Although the work done so far is only the first step towards using administrative registers for statistical purposes, a number of important lessons have already been learned that might be useful for other statistical agencies at the same transitional stage. It is crucial to have a strong legal basis that will provide the national statistics agency with the right to access administrative data at unit level with identifiers, and with the right to link them with other administrative registers for statistical purposes. On the other hand, there must be a high level of coordination and cooperation with the register managers to improve the quality of information that is collected administratively and when introducing potential adjustments to the collection forms. In the potential files of interest, the records must be evaluated in terms of coverage, content, quality, and identifiers. Identifiers play a considerable role in linking information from various sources. A potential line of work in overcoming the problem of incompatible identifiers between administrative registers is the use of the address as a key connection.

8. References

Andersen, E. and H. Utne. 2011. "Censuses in a Register-Based Statistical System: Norwegian Experiences." Paper presented at the 58th World Statistics Congress ISI 2011, IP064.01, 21–26 August, Dublin, Ireland.

Ballano, C. 2008. "A Census of Population Based on an Administrative Register." Paper presented at the 24th International Methodology Symposium, Statistics Canada, 28–31 October, Ottawa, Canada.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer. 2010. "A Quality Framework for Statistics Based on Administrative Data Sources Using the Example of the Austrian Census 2011." *Austrian Journal of Statistics* 39: 299–308.

Borchsenius, L. 2000. "From a Conventional to a Register-Based Census of Population." Paper presented at the INSEE/Eurostat Seminar on the Censuses after 2001, 20–21 November, Paris, France.

Conti, P., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28: 69–88.

Daas, P., S. Ossen, and J. Arends-Tóth. 2009. "Framework of Quality Assurance for Administrative Data Sources." Paper presented at the 57th World Statistics Congress ISI 2009, 16–22 August, Durban, South Africa.

Dolenc, D. 2010. "Quality Assessment in a Register-Based Census – Administrative Versus Statistical Concepts in the Case of Households." Paper presented at the European Conference on Quality in Official Statistics, 4–6 May, Helsinki, Finland.

Dugmore, K., P. Furness, B. Leventhal, and C. Moy. 2011. "Beyond the 2011 Census in the United Kingdom – With an International Perspective." *The Market Research Society* 53: 619–650.

Durr, J. and J. Dumais. 2002. "Redesign of the French Census of Population." *Survey Methodology* 28: 43–49.

Eppmann, H., S. Krügener, and J. Schäfer. 2006. "First German Register Based Census in 2011." *Allgemeines Statistisches Archiv* 90: 465–482. Doi: http://dx.doi.org/10.1007/s10182-006-0246-9.

Eurostat. 2011. *EU legislation on the 2011 Population and Housing Censuses – Explanatory Notes*. Luxembourg: Publications Office of the European Union.

Herman, E. 2008. "The American Community Survey: An Introduction to the Basics." *Government Information Quarterly* 25: 504–519. Doi: http://dx.doi.org/10.1016/j.giq.2007.08.006.

Houbiers, M. 2004. "Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands." *Journal of Official Statistics* 20: 55–75.

Instituto Nacional de Estatística (INE). 2010. *Programa de Acção para os Censos 2011*. Lisbon, Portugal.

Kish, L. 1986. "Complete Censuses and Sample." *Journal of Official Statistics* 2: 381–395.

Kish, L. 1990. "Rolling Samples and Censuses." *Survey Methodology* 16: 63–93.

Linder, F. 2004. "The Dutch Virtual Census 2001: A New Approach by Combining Administrative Registers and Household Sample Surveys." *Austrian Journal of Statistics* 33: 69–88.

Maldonado, A., D. Scheuregger, and K. Ziprik. 2010. "Setting Up the Central Register of Addresses and Buildings of the German Census 2011." Paper presented at the European Conference on Quality in Official Statistics, 4–6 May, Helsinki, Finland.

Mulry, M., S. Bean, D. Bauder, D. Wagner, T. Mule, and R. Petroni. 2006. "Evaluation of Estimates of Census Duplication Using Administrative Records Information." *Journal of Official Statistics* 22: 655–679.

Myrskylä, P. 1991. "Census by Questionnaire – Census by Registers and Administrative Records: The Experience of Finland." *Journal of Official Statistics* 7: 457–474.

Nordholt, E. 2005. "The Dutch Virtual Census 2001: A New Approach by Combining Different Sources." *Statistical Journal of the United Nations Commission for Europe* 22: 25–37.

Nordholt, E. and F. Linder. 2007. "Record Matching for Census Purposes in the Netherlands." *Statistical Journal of the IAOS* 24: 163–171.

Nordholt, E., S. Ossen, and P. Daas. 2011. "Research on the Quality of Registers to Make Data Decisions in the Dutch Virtual Census." Paper presented at the 58th World Statistics Congress ISI 2011, STS050.01, 21–26 August, Dublin, Ireland. Available at: http://2011.isiproceedings.org/Abstracts/STS050.html (accessed April 2012).

Ralphs, M. and P. Tutton. 2011. "Beyond 2011: International Models for Census Taking: Current Processes and Future Developments." Working Paper: Beyond 2011 Project - Office for National Statistics. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes—projects/beyond-2011/news/reports-and-publications/early-reports-and-research-papers/international-models-for-census-taking.pdf (accessed April 2012).

Redfern, P. 1986. "Which Countries Will Follow the Scandinavian Lead in Taking a Register-Based Census of Population?" *Journal of Official Statistics* 2: 415–424.

Redfern, P. 1989. "Population Registers: Some Administrative and Statistical Pros and Cons." *Journal of the Royal Statistical Society, Series A* 152: 1–41. Doi: http://dx.doi.org/10.2307/2982819.

Scheuren, F. 1999. "Administrative Records and Census Taking." *Survey Methodology* 25: 151–160.

Statistics Denmark. 1995. *Statistics on Persons in Denmark – A Register-Based Statistical System*. Luxembourg: Eurostat and Denmark Statistics.

Statistics Finland. 2004. *Use of Registers and Administrative Data Sources for Statistical Purposes – Best Practices of Statistics Finland*. Handbook 45, Helsinki, Finland.

Statistics Portugal. 2010. "Implementing a Register-Based Census in Portugal: Changing the Paradigm." Paper presented at Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, 10–11 May, The Hague, The Netherlands.

Szenzenstein, J. 2005. "The New Method of the Next German Population Census." *Statistical Journal of the United Nations Commission for Europe* 22: 59–71.

Tönder, J.-K. 2008. "The Register-Based Statistical System: Preconditions and Processes." Paper presented at the International Association for Official Statistics Conference, 14–18 October, Shanghai, China.

United Nations Economic Commission for Europe (UNECE). 2006. *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and*

Housing, Prepared in cooperation with the Statistical Office of the European Communities (Eurostat). New York and Geneva: United Nations Publications.

United Nations Economic Commission for Europe (UNECE). 2007. *Register-Based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics*. Geneva, Switzerland: United Nations Publications.

United Nations Economic Commission for Europe (UNECE). 2013. "Census Methodology: Key Results of the UNECE Survey on National Census Practices, and First Proposals about the CES Recommendations for the 2020 Census Round." Paper presented at the Conference of European Statisticians – Group of Experts on Population and Housing Censuses, 30 September – 3 October, Geneva, Switzerland. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/3_E_x_15_Aug_WEB_revised_map.pdf (accessed September 2014).

United Nations Statistic Division (UNSD). 2010. *Report on the Results of a Survey on Census Methods Used by Countries in the 2010 Census Round*. Working Paper: UNSD/DSSB/1. Available at: http://unstats.un.org/unsd/demographic/sources/census/2010_phc/docs/ReportOnSurveyFor2010Census.pdf (accessed April 2012).

Valente, P. 2010a. "Main Results of the UNECE/UNSD Survey on the 2010/2011 Round of Census in the UNECE Region". Paper presented at the Working Group on Demography and Census - Eurostat, 19–20 April, Luxembourg.

Valente, P. 2010b. "Census Taking in Europe: How Are Populations Counted in 2010?" *Population & Societies* 467: 1–4.

Valente, P. 2011. "Innovative Approaches to Census-Taking: Overview of the 2011 Census Round in Europe." Paper presented at the conference "Statistics in the 150 years from Italian Unification", June 8–10, Bologna, Italy.

Wallgren, A. and B. Wallgren. 2007. *Register-Based Statistics – Administrative Data for Statistical Purposes*. Chichester, UK: John Wiley & Sons.

Winkler, W. 2011. "Machine Learning and Record Linkage." Paper presented at the 58th World Statistics Congress ISI 2011, IPS057.01, 21–26 August, Dublin, Ireland. Available at: http://2011.isiproceedings.org/papers/450070.pdf (accessed January 2016).

Woods, S. 2009. "Evaluating Population Estimates in the United States: Counting the Population Between the Censuses." *Government Information Quarterly* 26: 144–147.

Zanutto, E. and A. Zaslavsky. 2002. "Using Administrative Records to Improve Small Area Estimation: An Example from the U.S. Decennial Census." *Journal of Official Statistics* 18: 559–576.

Zhang, L. 2011. "A United-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics* 27: 415–432.

Zhang, L. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received April 2012
Revised September 2014
Accepted September 2015

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 113–127, http://dx.doi.org/10.1515/JOS-2016-0005

Constructing Synthetic Samples

Hua Dong[1] *and Glen Meeden*[2]

We consider the problem of constructing a synthetic sample from a population of interest which cannot be sampled from but for which the population means of some of its variables are known. In addition, we assume that we have in hand samples from two similar populations. Using the known population means, we will select subsamples from the samples of the other two populations which we will then combine to construct the synthetic sample. The synthetic sample is obtained by solving an optimization problem, where the known population means, are used as constraints. The optimization is achieved through an adaptive random search algorithm. Simulation studies are presented to demonstrate the effectiveness of our approach. We observe that on average, such synthetic samples behave very much like actual samples from the population of interest. As an application we consider constructing a one-percent synthetic sample for the missing 1890 decennial sample of the United States.

Key words: Sample survey; missing data; synthetic samples.

1. Introduction

The Minnesota Population Center (MPC) is an interdepartmental demography research group at the University of Minnesota. One major goal of the MPC is to create databases that can be utilized in the study of economic and social behavior. The Center has developed the Integrated Public Use Microdata Series (IPUMS-USA), which is available online and which consists, in part, of high-precision one-percent samples of the American population drawn from fifteen decennial federal censuses. A sample is composed of microdata consisting of a record for each person. These records are in turn organized into households, making it possible to study the characteristics of people in the context of their families or other coresidents. Unfortunately the complete records for the 1890 census were destroyed and now only certain summary statistics are available. For example, the family incomes for each particular family are missing but the average 1890 family income is known for many small regions of the country. Hence the Center now does not have a one-percent sample based on the complete 1890 census. In this article we will present a method that will allow a synthetic sample to be created for 1890 using the partial information from 1890 and the samples from 1880 and 1900.

Since overall the 1890 US population should not be that different from the 1880 and 1900 populations, it should be possible to construct a synthetic one-percent sample for 1890 using the one-percent samples from the 1880 and 1900 populations. The records in the synthetic sample should be chosen in such a way that their summary statistics closely

[1] Gilead Sciences, Inc. Foster City, CA 94404, U.S.A. Email: hdong@gilead.com
[2] School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A. Email: glen@stat.umn.edu

match the partial information for 1890. To accomplish this, we define a function that measures just how closely a possible synthetic sample matches the known population means. Since there will be many possible synthetic samples that nearly achieve the minimum of this function, our goal will not be to find an optimal synthetic sample. Instead we will be looking for one which is nearly optimal. Before considering this problem, however, we will first consider some simpler problems and present simulation results that demonstrate that our approach works well in these cases. More information about the MPC can be found at https://www.pop.umn.edu/.

From one point of view, the lack of a one-percent sample from the 1890 US population can be thought of as a massive missing-data problem where the entire sample is missing. Creating a synthetic sample is impossible unless there is some additional information that can be used, and we believe that this is indeed the case here. In the following, we will consider simpler versions of this problem and present simulation results which show that our approach can work. These simulations might suggest that our approach could be helpful in more standard missing-data problems where just some of the sample is missing. Here, however, our focus will be on the problem of creating a synthetic one-percent sample for the 1890 census.

In Section 2 we introduce a simple version of our problem. In Section 3 we propose an adaptive random search algorithm that will find a nearly optimal synthetic sample. Given an objective function defined over a large space, this technique is used to locate a point in the space whose value given by the objective function is very close to the global optimum of the objective function. In Section 4 we present simulations which show that our method works well for some simple versions of our problem. In Section 5 we use our algorithm on census data from 1900 and 1920 and partial information from 1910 to produce synthetic samples for 1910. If our approach produces good synthetic samples for this situation, then we believe it should produce a good synthetic sample for 1890 when using the 1880 and 1900 census data. Section 6 contains some final remarks.

2. A Simple Problem

Assume that there are three populations, Population 1, Population 2, and Population 3, and we believe that in some sense Population 2 is the "average" of the other two. (For our problem, the three populations can be thought as the records for 1880, 1890, and 1900 respectively.) Attached to each unit in the populations there is a pair of variables, say, X and Z. We suppose that in the three populations X and Z are related, but we make no model assumptions about this relationship. We do assume however that the mean of Z is known for the second population, that we have independent random samples from the first and third populations, and that for each unit in the samples the values of both X and Z are observed. A simple version of our problem is to use this limited information about the second population and the samples from the other two populations to construct a synthetic sample that is formed by taking elements from the other two samples and that will behave like an actual sample from the second population.

More formally, for $i = 1$ and 3 let $z_i = (z_{i,1}, \ldots, z_{i,n})$ be the observed values of Z in the random sample from population i where $n = 2m$. These will be considered fixed in what follows. If $s_1 = (i_1, \ldots, i_m)$ and $s_3 = (j_1, \ldots, j_m)$ where $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and

$1 \leq j_1 < j_2 < \cdots < j_m \leq n$ we denote the two possible subsamples of size m by z_{s_1} and z_{s_3} and denote the synthetic sample of size n formed by their union as

$$z_{s_1,s_3} = (z_{s_1}, z_{s_3}) = (z_{1,i_1}, \ldots, z_{1,i_m}, z_{3,j_1}, \ldots, z_{3,j_m}) = (z_{s_1,s_3,1}, \ldots, z_{s_1,s_3,n})$$

Finally, let μ_2 be the known mean of Z for the second population.

We need a function to measure how good the synthetic sample based on s_1 and s_3 actually is. For example, suppose that the sample mean of z_{s_1,s_3} is equal to μ_2; then we consider this to be an optimal solution for our problem. Although in theory there can be more than one such optimal solution, in practice there will almost never be even one synthetic sample that is optimal in this sense.

Let $p = (p_1, \ldots, p_n)$ be a probability vector belonging to Γ, the $n - 1$ dimensional simplex, and let

$$\Gamma_{\mu_2}(z_{s_1,s_3}) = \left\{ p : p \in \Gamma \text{ and } \sum_{i=1}^{n} p_i z_{s_1,s_3,i} = \mu_2 \right\}$$

This is the set of all probability vectors on z_{s_1,s_3} whose mean is equal to μ_2.

Let

$$h(p) = \sum_{i=1}^{n} (p_i - 1/n)^2 \quad \text{and} \tag{1}$$

$$p_{s_1,s_3} = \arg\min \{h(p) : p \in \Gamma_{\mu_2}(z_{s_1,s_3})\} \tag{2}$$

Then $h(p_{s_1,s_3})$ is our measure of how good z_{s_1,s_3} is as a synthetic sample for the second population. Given two possible synthetic samples, we will prefer the one that yields the smaller value of this function. So an optimal solution for our problem is any choice of s_1 and s_3 that gives the minimum value of $h(p_{s_1,s_3})$ over all possible synthetic samples. Our approach involves two steps. First, for a given s_1 and s_3, we need to find p_{s_1,s_3}. The second step involves searching for an s_1 and s_3 that minimize $h(p_{s_1,s_3})$.

Now for fixed s_1 and s_3, finding the value $h(p_{s_1,s_3})$ is just a standard quadratic programming problem and many software packages will have a function that will find a solution. That said, we do not know how to find explicitly the choices of s_1^* and s_3^*, which minimize $h(p_{s_1,s_3})$ over all possible synthetic samples. Instead we will conduct a random search over this space to find an approximate solution for our problem. There are $\binom{2m}{m}^2$ possible choices for s_1 and s_3, so one possibility would be to just randomly select a large number of choices for s_1 and s_3 and use the one that gives the best answer. But as m increases, the space we are searching over can become quite large and there are better search algorithms than random sampling. In the next section, we will explain our adaptive random search algorithm that seems to give sensible answers to our problem.

Finally, we note that we can include constraints on more than one variable. In particular, we could have more than one constraint involving the same variable. For example, if the mean and variance of Z were known, we could add a second constraint using its second moment.

3. The Algorithm

As we noted in the previous section, we cannot find explicitly s_1^* and s_3^*, a solution for our desired problem. On the other hand, even though an optimal solution must exist, there will usually be many other solutions that are almost as good. Our goal is not to find an optimal solution but to find just one of the possibly many synthetic samples that are nearly optimal.

To do this we will carry out a random search in the space of all possible synthetic samples. As we just noted above, one possibility would be to select at random a large number of values for (s_1,s_3) and keep the one which gives the smallest value of $h(p_{s_1,s_3})$. This is not very efficient, however, and better methods are available. One approach is to pick a starting point at random and then select at random a second point that is close to it. If the value of the function $h(p_{s_1,s_3})$ is smaller than its value at the first point, then we should move to this new point. If it is not, then we can pick another point at random from the neighborhood of the first point and repeat the process. If our function has a global minimum and no local minimums, we will eventually arrive in the neighborhood of the minimum. If there are local minimums, however, then we could get stuck at one of those points and never reach the neighborhood of the global minimum. A way to avoid this is to sometimes allow a move to a point with a larger $h(p_{s_1,s_3})$ value with positive probability. This probability should depend on both the relative sizes of the two values of the $h(p_{s_1,s_3})$s and the point we are at in the search process.

More formally, suppose we are in step l of our search, where (s_1^l, s_3^l) is our current state and we are considering moving to a new state or point in the space of synthetic samples, say (s_1^{l+1}, s_3^{l+1}). The first thing to note is that in the long run, rather than picking the new point at random, it is more efficient to pick one that is close by the current state. In our case, we will pick either s_1^l or s_3^l at random and then pick one of its entries at random and replace it by a new member, selected at random, from the appropriate full sample. Once we have determined (s_1^{l+1}, s_3^{l+1}), we can check if

$$h(p_{s_1^{l+1},s_3^{l+1}}) < h(p_{s_1^l,s_3^l}) \tag{3}$$

If this is the case then we should move to the new state. If the converse is true, then sometimes we will still want to move to the new state. This will allow us to escape from a point in the space which is a local minimum. For example, if the above equation is false then at step l one could move to the new synthetic sample with probability θ where

$$\theta = \frac{h(p_{s_1^l,s_3^l})}{h(p_{s_1^{l+1},s_3^{l+1}})} \frac{t}{a+l} \tag{4}$$

where $0 < t \leq a$ are specified constants. Note that this makes it less likely that we will move to a worse synthetic sample after lots of steps than earlier in the process. This makes sense, since we are more likely to be close to the optimal solution after many steps than when we were near the beginning of the process. We continue this process for a fixed, large number of steps and then stop. It is important to note that the "best" synthetic sample in the entire sequence need not be the state we were in when we stopped. It could have occurred much earlier and we just moved away from it. In fact, this is what usually happens in our problem.

The form of the function for θ in Equation 4 is just one of many that can be used in practice but it seemed to work well in our problem. This algorithm is just a special case of what is known as an adaptive random search. These methods have been used in a variety of problems for more than 50 years.

4. Some Simulation Examples

We conducted some simulation studies to see how our approach could work in practice.

4.1. First Example

We began by constructing three similar populations where we expect our approach to work well. Attached to each unit there are the values of two continuous variables and of two binary variables. We will denote these variables by U, V, X, and Y. The variable U will be a random sample from a gamma distribution with shape parameter γ and scale parameter one. The variable V will be a random sample from a gamma distribution with shape parameter λ and scale parameter one. The variable Y will be a random sample from a Bernoulli distribution where θ is the probability of observing a one. These variables will be independent. The final variable, X, will be constructed using logistic regression with the variable V. For a unit for which $V = v$ let $p(v)$ be the probability that its X variable has the value one. Then for our model

$$log(p(v)/(1 - p(v))) = \beta v$$

Using this model we generated three populations, each with 4,000 units. The parameter values for the three populations are given in Table 1. Note that the parameter values for the second population are the average of the other two in all cases. In addition, for each variable their distributions across the three populations are quite similar. In the second population the correlation between V and X was 0.18.

The first four rows of Table 2 give the results of 1,200 random samples, each of size 40, taken from the second population, where the population mean of each variable was estimated. For each variable the table gives the average value of the sample mean, its average absolute error, the average lower bound and average length of the usual 95% confidence interval and its frequency of coverage. The next four rows give the results when synthetic samples were constructed assuming that the true mean of V in the second population was known. These synthetic samples were also of size 40 and used 20 observations each from samples of size 40 taken from the other two populations. Note that the two results are very similar except that the confidence intervals for the synthetic

Table 1. Parameter values used to generate the three populations with four variables for the first example in Section 4

Population	γ	λ	β	θ
1	6	7	0.10	0.4
2	5	8	0.15	0.5
3	4	9	0.20	0.6

Table 2. Comparing the results from 1,200 samples of size 40 from Population 2 to 1,200 synthetic samples formed by combining samples from Populations 1 and 3 constraining on knowing the population mean of V for the first example in Section 4

	Mean	absErr	lowBd	Length	Coverage Rate
Variable	When sampling from the actual population				
U	5.02	0.29	4.34	1.37	0.932
V	7.99	0.35	7.12	1.74	0.950
X	0.75	0.056	0.62	0.27	0.943
Y	0.51	0.063	0.35	0.31	0.947
	Using synthetic samples				
U	5.08	0.26	4.38	1.40	0.962
V	7.98	0.0	7.05	1.86	1
X	0.74	0.052	0.60	0.27	0.968
Y	0.49	0.061	0.34	0.31	0.955

samples always contain the true mean of *V*. This must be the case because of the way they were formed. The constraint guarantees this.

One might wonder how the synthetic samples do when estimating population quantiles. In Table 3 the true values of the five quantiles of *V* in the second population are given. The next two rows give the average values of their standard estimates along with their average absolute error for 1,200 samples of size 40. The next two sets of two rows give the same information for synthetic samples formed by constraining on the true means of *U* and *X* in

Table 3. Comparing the results for estimating five quantiles of variable V for the first example in Section 4 when sampling from the population and when using three different constraining variables to construct synthetic samples. The results are based on 1,200 samples of size 40

	0.10 quantile	0.25 quantile	0.50 quantile	0.75 quantile	0.90 quantile
True	4.61	5.92	7.65	9.64	11.77
	When sampling from the actual population				
Mean of est	4.78	6.02	7.68	9.62	11.58
absErr	0.44	0.41	0.41	0.52	0.84
	Using synthetic samples formed by constraining on the mean of *U*				
Mean of est	4.66	5.97	7.66	9.70	11.81
absErr	0.44	0.41	0.43	0.51	0.81
	Using synthetic samples formed by constraining on the mean of *X*				
Mean of est	4.62	5.91	7.64	9.73	11.84
absErr	0.42	0.37	0.43	0.54	0.79
	Using synthetic samples formed by constraining on the mean of *V*				
Mean of est	4.58	5.87	7.61	9.67	11.79
absErr	0.37	0.30	0.28	0.33	0.60

the second population. We see that these results are very similar to those found using actual samples from the second population. Finally, the last two rows of the table give the results for synthetic samples formed by constraining on the mean of V from the second population. We see that these results are significantly better than using actual samples from the second population. What is the explanation for this perhaps surprising result?

This happens because knowing the mean of V in the second population is a very important piece of information. This fact, along with samples from two very similar populations, allows us to construct synthetic samples that on average are better than random samples drawn from the actual population. This is not a common situation, but we believe that something like this could hold for the 1890 census. Next we will consider an example where our approach does not work as well.

4.2. *Second Example*

Perhaps it is not so surprising that we can find good synthetic samples when the three populations are very similar. Here we will consider another example where they are less similar and in particular where a mean of the middle population is not approximately equal to the average of the means of the other two. In this example we assume that each population has two continuous variables, say U and V, which are independent and of course independent across the three populations. Suppose the mean of U in the ith population is $\mu_{u,i}$ while the mean in the ith population of V is $\mu_{v,i}$. In our simulation the values of the $\mu_{u,i}$s were equal to 8, 10, and 12, for $i = 1, 2,$ and 3, while the corresponding values of the $\mu_{v,i}$s were 8, 9, and 12 respectively. All the distributions were normal with a common standard deviation equal to two. Each population contained 4,000 units and we constructed synthetic samples of size 60 for the second population using random samples of size 80 from the other two. Each synthetic sample contained 30 units from each of the other samples. We considered estimating the mean and the population quantiles of the variable U in the second population using synthetic samples based on various constraints.

The results for estimating the means are in Table 4. When constraining on the $E(V)$ our point estimate for $E(U)$ behaves just like the one based on samples from the actual population because $\mu_{u,2} = (\mu_{u,1} + \mu_{u,3})/2$. On the other hand, our point estimate for $E(V)$ when constraining on $E(U)$ performs poorly because $\mu_{v,2}$ is not the average of the other two means for V.

In addition, note that when constraining on $E(V)$ the confidence intervals for $E(U)$ are too long. In other words, even though our synthetic samples are centered properly they are too spread out. This happens despite the fact that the populations all have the same variance. So even though the average of the means for U for the first and third populations is equal to the mean of U for the second population, they are just too far apart to get good synthetic samples using just this one constraint. We can overcome this problem if we have more information about U for the second population. Suppose we know both its mean and variance; then we can constrain on both the first and second moments of U when selecting a synthetic sample. We did this in another simulation where we constrained on both $E(U)$ and $E(U^2)$ and we see from Table 4 that the length of the intervals, on average, are nearly the same as those based on random samples from U.

Table 4. Comparing the results for estimating E(U) and E(V) for the second example in Section 4 when sampling from the population and when constraining on moments of U and V. The results are based on 1,000 samples of size 60

Variable	Mean	absErr	lowBd	Length	Coverage Rate
	When sampling from the actual population				
U	9.97	0.20	9.47	1.00	0.949
V	9.01	0.20	8.52	0.99	0.950
	When constraining on $E(V)$				
U	10.02	0.19	9.31	1.42	0.998
V	9.01	0.0	8.29	1.45	1
	When constraining on $E(U)$				
U	9.98	0.0	9.27	1.42	1
V	10.07	1.06	9.34	1.45	0.11
	When constraining on $E(U)$ and $E(U^2)$				
U	9.98	0	9.48	1.00	1
V	10.11	1.10	9.35	1.51	0.068
	When constraining on $E(U)$, $E(U^2)$, $E(V)$, and $E(V^2)$				
U	9.99	0.01	9.49	1.01	1
V	9.03	0.02	8.52	1.01	1

To explore this further, we next considered the results for estimating the quantiles of U for the second population. We see from Table 5 that when constraining on either $E(U)$ or $E(V)$ our synthetic samples tend to underestimate the 0.10 quantile and overestimate the 0.90 quantile. That is, our synthetic samples are too spread out. However, when our constraints include both $E(U)$ and $E(U^2)$, our estimates based on synthetic samples perform better than random samples from the actual population. Although we have not included the simulation results, the story is the same for estimating the quantiles of V.

4.3. Third Example

We produced another example where each population consists of samples from three independent normal random variables, say U, V, and W. In all cases their standard deviations were 1.5. The means of U across the three populations were 8, 10, and 12 respectively, while for V they were 8, 9, and 12 and for W 8, 11, and 12. Each population contained 4,000 units. Then we took 1,000 samples of size 120 from the first and third populations to construct a synthetic sample for the middle population of size 40 by using 20 units each from the two samples. We constructed synthetic samples where for each variable their first two sample moments agreed with the first two population moments for the middle population. For each sample we estimated the 0.10, 0.25, 0.50, 0.75, and 0.90 population quantiles by their corresponding sample quantiles. Both the real samples and synthetic samples were approximately unbiased. Averaged over all samples and all

Table 5. Comparing the results for estimating five quantiles of variable U for the second example in Section 4 when sampling from the population and when constraining on moments of U and V. The results are based on 1,000 samples of size 60

	0.10 quantile	0.25 quantile	0.50 quantile	0.75 quantile	0.90 quantile
True	7.43	8.66	10.00	11.31	12.53
	When sampling from the actual population				
Mean of est	7.49	8.68	10.00	11.28	12.45
absErr	0.35	0.28	0.24	0.29	0.35
	When constraining on $E(V)$				
Mean of est	6.36	7.83	10.05	12.08	13.61
absErr	1.07	0.85	0.24	0.77	1.08
	When constraining on $E(U)$				
Mean of est	6.40	7.87	9.98	12.02	13.61
absErr	1.04	0.79	0.23	0.71	1.08
	When constraining on $E(U)$ and $E(U^2)$				
Mean of est	7.49	8.68	10.00	11.28	12.33
absErr	0.23	0.20	0.14	0.11	0.28
	When constraining on $E(U)$, $E(U^2)$, $E(V)$, and $E(V^2)$				
Mean of est	7.55	8.36	9.85	11.47	12.70
absErr	0.18	0.31	0.17	0.17	0.22

quantiles, the average absolute error for the real samples was 0.28 and for the synthetic samples 0.17. We repeated this example but now using a standard deviation of 2 instead of 1.5. In this case, averaged over all samples and all quantiles the average absolute error for the real samples was 0.28 and for the synthetic samples it was 0.20. So in both cases the synthetic samples seem to give a good picture of the unsampled populations. Once again, this shows that one can construct good synthetic samples from samples of similar populations for a population for which some true population parameters are known. How good the synthetic samples will actually be depends on how similar the populations are and how much is known about the middle population.

4.4. Fourth Example

So far we have seen that our method seems to work well when the three populations are quite similar and we are estimating means and quantiles. It is natural to wonder how our method will work if we are interested in estimating more complicated population parameters, say a regression coefficient.

Consider three populations, each of which consists of two variables X and Y. Let μ_i denote the mean of X in the ith population. In population i, X is normally distributed with mean μ_i and standard deviation 5. The distribution of Y given $X = x$ is normal with mean $50 + \beta_i x$ and standard deviation 15. All three populations will contain 4,000 units.

In the first example we let $\mu_i = 200$, 205, and 210 for $i = 1$, 2, and 3 and set β_i to be equal to 2 for all three populations. For the middle population the true value of the regression parameter was 1.99 and the correlation between X and Y was 0.58. We then took 500 random samples of size 120 from the first and third populations and for each pair of random samples found a synthetic sample of size 60 by selecting 30 units each from the two random samples where we assumed the population means of X and Y were known for the middle population. For these 500 synthetic samples, we found that the average value of their estimates for β_2 was 1.95 with average absolute error of 0.23. The average length of their 95% confidence interval was 0.53 with a frequency of containing the true value equal to 0.928. The corresponding values for 500 random samples from the middle population were 2.01, 0.31, 0.78, and 0.930. So in this example the synthetic samples perform very well.

In a second example we set the three μ_is equal to 200 but let the three β_is be equal to 2.00, 2.15, and 2.30 for the three populations. For this case with 500 synthetic samples formed as in the previous paragraph, we found that the average value of our estimates was 2.21 with an average absolute error of 0.70. The average length of the 95% confidence intervals was 1.71 with a frequency of containing the true value equal to 0.924. The corresponding values for 500 random samples from the middle population were 2.18, 0.33, 0.778, and 0.932. So here our synthetic samples are not doing so well. We believe this happens because in this second example the three populations are not quite as similar as those in the previous example. We find it interesting, however, that the confidence intervals based on the synthetic samples have approximately the correct coverage probability in both examples. In any case, it is clear from all our simulations that how well synthetic samples work depends not only on how similar the three populations are but also on what population parameters are being estimated.

4.5. *Behavior of the Algorithm*

Recall that our goal is not to find an optimal synthetic sample but just one among the large group of those who are nearly optimal. For the example in Subsection 4.3 where the standard deviation was 2, we ran our adaptive random search algorithm for 20,000 steps for each sample. We kept track of how many times it moved to a new state, the time it moved to the best state, our solution, and the time of the last move. For this example, on average, our chain moved to 96 new states, the last move occurred at step number 8,500 and our solution occurred at step number 8,055. The average of the minimum of the p_is in our solution was 0.024. Note that if our solution satisfied the constraints exactly, all the p_is would equal 1/40 $=$ 0.025. For the case where the standard deviation was 1.5, we ran our algorithm for 40,000 steps because there is more separation among the three populations. For this case, on average, our chain visited 174 states with the last move happening at step 19,537 and our solution occurring at step 15,533. The average of the minimum of the p_is in our solutions was 0.023.

Readers might have been questioning the need for using an adaptive random search algorithm and whether using random sampling for the searching could work just as well. For the above problem we took 100 random samples and for each sample we selected 20,000 possible synthetic samples at random. For each sample we found the value of the vector p_{s_1,s_3}, which is the solution to the problem given in Equation 2. We then found the synthetic sample, which minimized the function h in Equation 1 over all the random samples. Averaged over these 100 samples the average minimum value of p was 0.0076. We repeated this but now included 100,000 random samples in our search. In this case, the average of the minimum values of p was 0.0084. Finally, we used 400,000 random samples for our search and found that this average was 0.0102. So even taking 20 times the number of synthetic samples that we do using our method, random sampling cannot find any synthetic samples that are as nearly balanced as ours.

Clearly our solutions are not optimal, but they are good enough for the synthetic samples to be good representations of real samples because we are constructing synthetic samples from samples of two populations that are similar to the population of interest.

5. A Simulation Using Census Data

To look at the potential performance of the proposed method for the missing 1890 population problem, we tested the proposed method on some actual census data from nearby decades. We used data supplied from the MPC for one geographical area out of a total of 56 possible geographical areas. We had approximately 2.3% samples from 1900, 1910, and 1920, which we treated as the entire populations. Associated with each individual was a vector of possible values indicating gender, age, marital status and race. We then selected 100 random samples of size about 100 from the 1900, 1910, and 1920 populations. We assumed that the sample from 1910 is missing and only the population means of five constrained variables were known. The five constrained variables were "married males", "single males", "married females", "single females", and "Negroes". We used the population means of these five variables from our 1910 population as our mean constraints and samples from the 1900 and 1920 "populations" to construct synthetic samples which contained about 50 individuals each from 1900 and 1920.

Because individuals are members of households, when a person was selected to be in our sample we included everyone in their household as well. Our samples always included at least 100 individuals. Our synthetic samples also always included at least 100 individuals. At each step of the search it was possible that we would need to remove more than one household to reduce the size of the current synthetic sample to be less than 100. By the same token, we might also need to add more than one household to ensure the number of individuals in the next synthetic sample would be at least 100. So a possible synthetic sample need not contain exactly 50 observations from 1900 and 1920 respectively.

To see what happens in this case, we constructed 100 synthetic samples using samples from 1900 and 1920 and the true 1910 population means as constraints. The results are given in Table 6.

To gain a better understanding of how the synthetic samples work we did another simulation where instead of constraining on the population means of the five variables we used sample information. That is, each time we took a sample from the 1910 population as well and used the sample means of our five constraining variables as the constraints when constructing a synthetic sample for 1910 from the samples from 1900 and 1920.

Table 6. The results for the synthetic samples for the 1910 population when the true population means are used as constraints

Variable	Mean	absErr	SD	Margin of error	Coverage rate
Married males	0.178	0.006	0.384	0.074	1.000
Divorced males	0.002	0.002	0.016	0.003	0.160
Widowed males	0.013	0.009	0.097	0.019	0.780
Single males	0.308	0.006	0.464	0.089	1.000
Negroes	0.352	0.004	0.480	0.092	1.000
Mulattoes	0.035	0.048	0.133	0.025	0.370
Married females	0.181	0.006	0.387	0.074	1.000
Divorced females	0.002	0.004	0.018	0.004	0.180
Widowed females	0.037	0.009	0.187	0.036	0.950
Single females	0.280	0.007	0.451	0.087	1.000
Foreign born	0.007	0.010	0.047	0.009	0.340
Age	23.093	1.451	18.628	3.571	0.950

Table 7 contains the results and is based on 100 samples. For comparison, we also calculated the estimates using the actual 100 samples from the 1910 population.

Note that the point estimates and the length of the confidence intervals based on the synthetic samples are very similar in the two tables. The intervals in Table 6 have better coverage rates, however. The better results occur because we are using better information, true population means, as our constraints.

For our purposes the more important fact is that the results for the synthetic samples are very similar to the results for the real samples in Table 6. This happens because the difference between the 1910 means and the average of the 1900 and 1920 means is quite small for most variables we considered. Because of the small size of our samples, it is not surprising that, especially for the rarer categories, the coverage rates of the confidence intervals can fall short of 95%. Moreover, we would expect the synthetic sample to perform poorly for a category whose 1910 mean is different from the average of its means from 1900 and 1920. For example, the coverage rate of the confidence intervals for “mulattoes” from the true 1910 sample is 0.53, which is much higher than 0.31, the coverage rate for the synthetic samples from 1900 and 1920. We believe that this stems from fact that the population proportion of “mulattoes” in 1910 is about 0.068, which is much higher than 0.028 which is the average of 1900 and 1920 population proportions. Note also that the margin of errors for the actual and synthetic samples are very similar. Because of the similarity of the three populations and the fact that the majority of the variables are binary, we see that just constraining on first moments is enough to obtain intervals with about the right length.

In our simulations, using the adaptive random search method based on Equation 4, we stopped the iterations after 5,000 steps. When trying to find one particularly good synthetic sample, there is no reason to stop after a particular number of steps. We did it here to make the running of a set of simulations easier. Since for the 1890 problem we are only interested in creating one sample, running the algorithm a long time is not a problem. However, it could take some experimentation to come up with a good choice for the values of t and a in Equation 4, as the number of variables used as constraints varies.

Table 7. A comparison of actual and synthetic samples for the census data when constraints based on sample information is used

Sample	Variable	Mean	absErr	SD	Margin of error	Coverage rate
1910	Married males	0.156	0.030	0.363	0.070	0.920
1910	Divorced males	0.001	0.002	0.008	0.002	0.080
1910	Widowed males	0.010	0.010	0.078	0.015	0.640
1910	Single males	0.333	0.042	0.471	0.091	0.950
1910	Negroes	0.325	0.101	0.450	0.087	0.510
1910	Mulattoes	0.077	0.052	0.224	0.043	0.530
1910	Married females	0.157	0.030	0.364	0.070	0.910
1910	Divorced females	0.003	0.004	0.029	0.006	0.280
1910	Widowed females	0.026	0.016	0.149	0.029	0.740
1910	Single females	0.314	0.050	0.463	0.089	0.870
1910	Foreign born	0.010	0.012	0.057	0.011	0.350
1910	Age	21.352	2.643	17.416	3.364	0.650
synthetic	Married males	0.157	0.028	0.364	0.070	0.940
synthetic	Divorced males	0.001	0.002	0.008	0.002	0.080
synthetic	Widowed males	0.009	0.009	0.074	0.014	0.620
synthetic	Single males	0.331	0.037	0.471	0.090	0.980
synthetic	Negroes	0.326	0.099	0.452	0.087	0.530
synthetic	Mulattoes	0.026	0.050	0.109	0.021	0.310
synthetic	Married females	0.159	0.027	0.366	0.070	0.960
synthetic	Divorced females	0.001	0.003	0.014	0.003	0.140
synthetic	Widowed females	0.027	0.016	0.151	0.029	0.760
synthetic	Single females	0.315	0.048	0.464	0.089	0.900
synthetic	Foreign born	0.010	0.011	0.063	0.012	0.410
synthetic	Age	21.614	2.170	17.418	3.336	0.810

Another approach to the 1890 census problem could be to try to use the information from the 1880 and 1900 censuses to create a model for the 1890 population that would then be used to generate a sensible one-percent census for 1890. Although such an approach could work, building a model for the entire US population would be a big problem. We believe, however, that the approach used here is simpler, and it effectively uses the information available in the 1880 and 1900 censuses in a simple and straightforward manner that bypasses the difficult problem of trying to construct a sensible model.

On the other hand, when constructing a one-percent sample for 1890 for a particular geographic area, historical information should be used when selecting the variables to constrain upon. These variables could depend on which area of the country you are considering. For rarer groups, you could make sure that each synthetic sample contains about the right proportion of individuals of that type. For example, if a family with a foreign-born individual is removed then it must be replaced by another family containing a foreign-born individual. If a proposed synthetic sample does not have approximately the correct mean for some variable not included in the constraining set, then one can always add this variable to the constraint set and find a new synthetic sample. Since a synthetic sample for the whole country will be made up of a collection of synthetic samples for a large number of many small geographic areas, the approach given here should be able to construct a good synthetic sample for the 1890 population.

6. Final Remarks

Here we have considered the problem of constructing a synthetic sample from a population for which we have limited information. We proposed a novel approach that assumes the existence of two known populations which taken together are a good approximation to the missing population. We have seen in some cases that a synthetic sample can be constructed and used as a substitute for a missing sample and inferences based on it are as good as those based on the actual sample. In particular, we saw that to get synthetic samples that do a good job of estimating the quantiles of a variable one can constrain on the first two moments of the variable. To obtain the synthetic sample, we used an adaptive random search algorithm to solve an optimization problem which incorporates the available limited information about the population of interest. Simulations demonstrated the good performance of our approach for some small sample sizes.

As we have pointed out, creating a synthetic one-percent sample for the 1890 census is an extreme missing-data problem, and as far as we know this problem has never been considered in the literature. Although this is perhaps stating the obvious, we were not interested in combining or merging two data sets, a problem which has often been discussed in the literature (Kadane 2001). On the other hand, synthetic data has been considered in several contexts. It has been recommended to replace missing or censored observations with imputed or synthetic observations. In some such cases auxiliary information is used to model the missing observations. In the survey-sampling context, after a sample has been selected Hidiroglou and Laniel (2001) considered constructing synthetic variables at the estimation stage. In situations where confidentiality is an issue, Fienberg et al. (1998) considered constructing synthetic samples as part of a disclosure-avoidance methodology, but they were modifying existing samples rather than constructing new ones. Reiter (2002), Reiter (2005), and Drechsler and Reiter (2012) recommended constructing many synthetic samples and then using multiple imputation to make inferences. It was argued that valid inferences could still be made using such synthetic data. Multiple imputation is not an option for the MPC since the goal is to create a one-percent sample for the 1890 census. In a situation closer to our problem, Kohnen and Reiter (2009) considered combining information from two populations, but again they use multiple imputation to construct many synthetic samples. Meeden (2000) gives an approach to the standard missing-data problem involving constraints that is closer in spirit to what we are doing here. There, after one set of values are imputed for the missing observations, the observed and imputed values are then adjusted so that confidence intervals based on this adjusted sample will have the correct frequentist coverage probability under repeated sampling.

Another possible application of our methods is to create a synthetic sample for a population using samples from similar populations and constraints based on partial information from a sample taken from the population of interest. In one case here, we saw that such synthetic samples worked well. We have carried out other simulation studies, not included here, and observed that if the three populations are not too different such synthetic samples behave very much like actual samples from the population. Although real data are always preferred, it seems clear to us that in some cases inferences based on synthetic data can perform almost as well as inferences based on actual data.

7. References

Drechsler, J. and J. Reiter. 2012. "Combining Synthetic Data with Subsampling to Create Public Use Microdata Files for Large Scale Surveys." *Survey Methodology* 38: 73–79.

Fienberg, S., U. Makov and R. Steele. 1998. "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data." *Journal of Official Statistics* 14: 485–502.

Hidiroglou, M. and N. Laniel. 2001. "Sampling and Estimation Issues for Annual and Subannual Canadian Business Surveys." *International Statistical Review* 69: 487–504. Doi: http://dx.doi.org/10.1111/j.1751-5823.2001.tb00471.x.

Kadane, J. 2001. "Some Statistical Problems in Merging Datasets." *Journal of Official Statistics* 17: 423–433.

Kohnen, C. and J. Reiter. 2009. "Multiple Imputation for Combining Confidential Data Owned by Two Agencies." *Journal of the Royal Statistical Society, Series A* 172: 511–528. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00574.x.

Meeden, G. 2000. "A Decision Theoretic Approach to Imputation in Finite Population Sampling." *Journal of the American Statistical Association* 95: 586–595. Doi: http://dx.doi.org/10.1080/01621459.2000.10474234.

Reiter, J. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18: 531–543.

Reiter, J. 2005. "Releasing Multiply Imputed, Synthetic Public Use Micro-Data: An Illustration and Empirical Study." *Journal of the Royal Statistical Society, Series A* 168: 185–205. Doi: http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

Received January 2013
Revised February 2015
Accepted November 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 129–145, http://dx.doi.org/10.1515/JOS-2016-0006

A Discussion of Weighting Procedures for Unit Nonresponse

David Haziza[1] *and Éric Lesage*[2]

Weighting procedures are commonly applied in surveys to compensate for nonsampling errors such as nonresponse errors and coverage errors. Two types of weight-adjustment procedures are commonly used in the context of unit nonresponse: (i) nonresponse propensity weighting followed by calibration, also known as the two-step approach and (ii) nonresponse calibration weighting, also known as the one-step approach. In this article, we discuss both approaches and warn against the potential pitfalls of the one-step procedure. Results from a simulation study, evaluating the properties of several point estimators, are presented.

Key words: Calibration; nonresponse bias; one-step approach; propensity-score adjusted estimator; two-step approach; unit nonresponse.

1. Introduction

Weighting procedures are commonly applied in surveys to compensate for nonsampling errors such as nonresponse errors and coverage errors. Brick (2013) provides an excellent overview of weighting in the presence of unit nonresponse; see also Kalton and Flores-Cervantes (2003). Two types of weight-adjustment procedures are commonly used in the context of unit nonresponse: (i) nonresponse propensity weighting followed by calibration, also known as the two-step approach and (ii) nonresponse calibration weighting, also known as the one-step approach. In this article, our focus is to warn against the potential pitfalls of the one-step procedure.

The two-step approach consists of adjusting the weights in two distinct steps: the basic (design) weights of respondents are first multiplied by a nonresponse adjustment factor, which is defined as the inverse of the estimated response probability. The adjusted weights are further modified so that survey-weighted estimates agree with known population totals. In the first step, survey statisticians aim at reducing the nonresponse bias, which may be appreciable when respondents and nonrespondents are different with respect to the survey variables. Whether or not one will succeed in achieving an efficient bias reduction depends on the availability of powerful auxiliary information (Särndal and Lundström 2005), which is a set of variables available for both respondents and nonrespondents. In the second step, some form of calibration (e.g., poststratification) is performed in order to ensure consistency between survey-weighted estimates and known population totals.

[1] Université de Montréal, C.P. 6128, Centre-ville Montreal Quebec H3C 3J7 Canada. Email: david.haziza@umontreal.ca
[2] INSEE, 18 bd Adolphe Pinard 75 014 Paris, France. Email: eric.lesage@insee.fr

Acknowledgments: The authors are grateful to an associate editor and three reviewers for their comments and suggestions, which substantially improved the quality of this article. David Haziza's research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

Calibration procedures require that the auxiliary variables (called calibration variables) are available for the respondents and that their population totals are known. In practice, the calibration variables are often specified by survey managers, who wish to ensure consistency with respect to some important variables (e.g., age and sex). Moreover, if the calibration variables are related to the characteristics of interest, the resulting calibration estimators tend to be more efficient than the noncalibrated ones.

The one-step approach pursues the same three goals as the two-step approach: reduce the nonresponse bias, ensure consistency between survey estimates and known population totals and, possibly, reduce the variance of point estimators. However, the weighting process is performed in a single step and does not require explicit estimation of the response probabilities.

In the absence of nonsampling errors, calibration consists of determining a set of calibrated (or final) weights as close as possible to the basic weights, while satisfying calibration constraints. A calibrated weight is expressed as the basic weight multiplied by a calibration adjustment factor, which depends on a calibration function. Commonly used calibration functions include the linear function, the exponential function, the truncated linear function and the logit function; see Section 2. Deville and Särndal (1992) showed that calibration estimators are asymptotically design consistent and that all the distance functions are asymptotically equivalent in the sense that they all lead to calibration estimators that are asymptotically equivalent to the calibration estimator based on the linear calibration function. The calibration function is usually chosen so that the distribution of the calibrated weights is "cosmetically attractive". For example, a problem that can be encountered with the linear function is the occurrence of negative weights, which can be prevented by using the exponential function that ensures positive weights. However, the latter may lead to extreme weights, which in turn may contribute to increase the instability of point estimators for characteristics of interest weakly correlated with the calibration variables. In this case, functions such as the truncated linear function or the logit function can be used in order to ensure that the calibration adjustment factors lie between prespecified lower and upper bounds.

How to choose the calibration function in the presence of unit nonresponse? In the case of the two-step approach, calibration is performed after the weights have been adjusted for nonresponse. As a result, the choice of the calibration function can be essentially made using the same criteria as in the complete response case. This is discussed further in Section 3. The situation is more intricate with the one-step approach, as different calibration functions may lead to calibration estimators with substantially different properties in terms of bias and mean square error. As a result, the choice of the calibration function is generally important when calibration is used for treating nonresponse. While the choice of calibration variables has been widely discussed in the literature (e.g., Särndal and Lundström 2005 and Särndal 2011), the issues of how to select an appropriate calibration function in the context of the one-step approach and the effect of function misspecification on the properties of the resulting estimators have not received a lot of attention. Two notable exceptions are Kott (2006) and Kott and Liao (2012). In this article, we argue that, even though the one-step approach does not use estimated response probabilities in the construction of point estimators explicitly, a wrong choice of the calibration function can have inadvertent and detrimental effects, even in the presence of high association between the auxiliary variables and the study variable. The matter deserves more careful attention than what it seems has hitherto been noticed in the literature; see Section 4. In Section 5, we show empirically that an inappropriate calibration function may lead to biased calibration estimators (sometimes exhibiting a bias larger than that of unadjusted estimators). This is especially true in the presence of quantitative auxiliary variables. The paper ends with a discussion in Section 6.

2. Calibration Weighting in the Complete Data Case

Let $U = \{1, 2, \ldots, N\}$ be a finite population consisting of N elements. Most surveys conducted by statistical agencies are multipurpose surveys, which are designed to provide statistics for a possibly large number of variables. For simplicity, we use the generic notation y to denote a characteristic of interest. In this paper, we are interested in estimating a population total $t_y = \sum_{k \in U} y_k$, where y_k denotes the k-th value of the characteristic of interest y, $k = 1, \ldots, N$. A sample s, of size n, is selected from U according to a given sampling design $p(s)$. Let π_k denote the first-order inclusion probability of unit k in the sample and $d_k = 1/\pi_k$ denote its design weight. Applying the basic weighting system, $\{d_k;\ k \in s\}$, to a y-variable leads to the well-known Horvitz-Thompson estimator

$$\hat{t}_{y\pi} = \sum_{k \in s} d_k y_k. \tag{1}$$

The estimator (1) is design unbiased for t_y regardless of the characteristic of interest y being estimated. That is, $E_p(\hat{t}_{y\pi}) = t_y$, where $E_p(\cdot)$ denotes the expectation with respect to the sampling design.

In practice, auxiliary information is often available at the estimation stage. Let $\mathbf{x}_i = (x_{1i}, \ldots, x_{Ji})^\top$ be a J-vector of auxiliary variables attached to unit i. We assume that the vector of population totals, $\mathbf{t}_\mathbf{x} = (t_{x_1}, \ldots, t_{x_J})^\top$, is known without error, where $t_{x_j} = \sum_{i \in U} x_{ji}$. While the basic weighting system ensures unbiasedness, that is, $E_p(\hat{\mathbf{t}}_{\mathbf{x}\pi}) = \mathbf{t}_\mathbf{x}$, it does not generally produce an exact estimate for each of the J auxiliary variable; that is, $\hat{\mathbf{t}}_{\mathbf{x}\pi} \neq \mathbf{t}_\mathbf{x}$, in general. To overcome the problem, we seek a calibrated weighting system $\{w_k;\ k \in s\}$ such that the weights w_k are "as close as possible" to the design weights d_k while satisfying the calibration constraints

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{x}.$$

The resulting calibrated weight w_k is given by

$$w_k = d_k F(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k), \tag{2}$$

where $F(\cdot)$ is a monotonic and twice-differentiable function such that $F(0) = 1$ and $F'(0) = 1$ and $\hat{\boldsymbol{\lambda}}$ is a J-vector of estimated coefficients (Deville and Särndal 1992). The weight w_k in (2) is the product of the design weight d_k and the calibration adjustment factor $F(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$. The calibration factor $F(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$ depends on (i) the calibration function $F(\cdot)$, (ii) the characteristics of unit k through $\mathbf{x}_k$ and (iii) the vector of estimated coefficients $\hat{\boldsymbol{\lambda}}$, which can be viewed as a measure of sample imbalance. Under mild regularity conditions, Deville and Särndal (1992) showed that $\hat{\boldsymbol{\lambda}} \rightarrow 0$ in probability as $n \rightarrow \infty$ and $N \rightarrow \infty$.

The resulting calibration estimator is

$$\hat{t}_C = \sum_{k \in s} w_k y_k. \tag{3}$$

Several calibration functions $F(\cdot)$ are used in practice, each corresponding to a particular calibration method. The most popular calibration methods are: (i) the linear method

$$F(u) = 1 + u; \tag{4}$$

(ii) the exponential method

$$F(u) = \exp(u); \tag{5}$$

(iii) the truncated linear method

$$F(u) = \begin{cases} 1 + u & L - 1 \leq u \leq M - 1 \\ M & u > M - 1 \\ L & u < L - 1, \end{cases} \tag{6}$$

where L and M are the prespecified lower and upper bounds, respectively; and (iv) the logit method

$$F(u) = \frac{L(M-1) + M(1-L)\exp(Au)}{M - 1 + (1-L)\exp(Au)}, \tag{7}$$

where

$$A = \frac{M - L}{(1-L)(M-1)}.$$

Assuming that the inverse of $\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k^\top$ exists, the linear method leads to a closed-form solution. In contrast, Methods (5)–(7) require some numerical methods that may fail to converge in some situations. However, the linear method may produce negative calibration adjustment factors, $F(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$, resulting in negative calibrated weights. On the other hand, the exponential method ensures that the calibration adjustment factors are positive, although some could be extreme. To avoid unduly large calibration adjustment factors, one can specify lower and upper bounds through the use of the truncated linear and logit methods. Deville and Särndal (1992) showed that the calibration estimator (3) is design consistent and approximately design unbiased for t_y regardless of the characteristic y being estimated and that all the calibration methods are asymptotically equivalent in the sense that they all lead to the calibration estimator based on the linear method.

We now discuss two important situations that are frequently encountered in practice. Let x_1 and x_2 be two categorical variables with J_1 and J_2 categories, respectively. The population U is then divided into $J_1 \times J_2$ cells. Let $N_{j_1 j_2}$ be the population count corresponding to the (j_1, j_2) cell, $j_1 = 1, \ldots, J_1$ and $j_2 = 1, \ldots, J_2$. Two cases may occur in practice: (i) the population counts $N_{j_1 j_2}$ are known. This case corresponds to a standard poststratification based on a vector of auxiliary information of size $J = J_1 \times J_2$. It is worth noting that, in this case, the choice of the calibration function $F(\cdot)$ is unimportant as all the calibration functions lead to the same calibrated weighting system $\{w_k;\ k \in s\}$. (ii) The individual cell counts $N_{j_1 j_2}$ are not known but the population margins $N_{j_1 \bullet} = \sum_{j_2=1}^{J_2} N_{j_1 j_2}$ and $N_{\bullet j_2} = \sum_{j_1=1}^{J_1} N_{j_1 j_2}$ are known, leading to a vector of auxiliary information of size $J = J_1 + J_2$. In this context, Deville et al. (1993) showed that the use of the exponential method (5) leads to the raking ratio estimator. Unlike case (i), different calibration functions generally lead to different calibrated weighting systems in case (ii). This discussion can be extended to more than two categorical variables. In this instance, case (ii) is often referred to as generalized raking procedures. We revisit both situations in Section 6 in the context of nonresponse adjustment.

3. The Two-Step Approach: Nonresponse Propensity Weighting Followed by Calibration

In the presence of unit nonresponse, the characteristics of interest are observed for a subset, s_r, of the original sample s. Let ϕ_k be the unknown response propensity attached to unit k. We assume that $\phi_k > 0$ for all k and that units respond independently of one another. We postulate the following nonresponse model

$$\phi_k = m(\mathbf{z}_k, \gamma), \tag{8}$$

where $m(\cdot)$ is a given function, $\mathbf{z}_k$ is a vector of auxiliary variables available for both respondents and nonrespondents and γ is a vector of unknown parameters. In this article, we assume that the $\mathbf{z}$-vector is correctly specified but not necessarily the functional form of (8). The choice of the $\mathbf{z}$-vector is discussed in Little and Vartivarian (2005).

In the first step, an estimate of ϕ_k is $\hat{\phi}_k = m(\mathbf{z}_k, \hat{\gamma})$, where $\hat{\gamma}$ is a suitable estimator of γ. The adjusted weight for nonresponse attached to unit k is defined as $\tilde{w}_k = d_k/\hat{\phi}_k$ for $k \in s_r$, leading to a weighting system adjusted for nonresponse, $\{\tilde{w}_k;\ k \in s_r\}$. The factor $\hat{\phi}_k^{-1}$ is often called the nonresponse adjustment factor for unit k. Applying the weighting system $\{\tilde{w}_k; k \in s_r\}$ to a characteristic of interest y leads to the Propensity-Score Adjusted (PSA) estimator of t_y (e.g., Lee 2006):

$$\hat{t}_{PSA} = \sum_{k \in s_r} d_k \hat{\phi}_k^{-1} y_k = \sum_{k \in s_r} \tilde{w}_k y_k. \tag{9}$$

The rationale behind this type of weighting procedure is similar in spirit to weighting for two-phase sampling.

Estimates of the ϕ_k's may be obtained through the use of a parametric model; for example, a logistic regression model as found in Ekholm and Laaksonen (1991). In the context of parametric nonresponse models, Kim and Kim (2007) showed that the PSA estimator (9) is asymptotically unbiased and consistent for t_y regardless of the characteristic y being estimated if (8) is correctly specified. However, parametric methods are rarely used in practice because some estimates $\hat{\phi}_k$ may be very small, leading to extreme nonresponse adjustment factors, ultimately resulting in highly dispersed weights $\tilde{w}_k$. Moreover, parametric methods are vulnerable to the misspecification of $m(\cdot)$.

In practice, nonparametric methods are preferred. A popular method, called the score method (Haziza and Beaumont 2007), consists of first obtaining preliminary estimated response probabilities $\tilde{\phi}_\mathrm{k}$ using a parametric model (e.g., the logistic regression model) and partitioning the sample into homogeneous weighting classes formed on the basis of the $\tilde{\phi}_k$'s. The basic weight of a respondent in a given class is then adjusted using the observed response rate within the same class (e.g., Little 1986; Eltinge and Yanaseh 1997). Other nonparametric methods include smoothing methods such as kernel and local polynomial methods (e.g., Giommi 1987; Da Silva and Opsomer 2006, 2009) and regression trees (e.g., Phipps and Toth 2012). Nonparametric methods are expected to provide some robustness if the form of $m(\cdot)$ is misspecified and protect (to some extent) against the noninclusion of predictors accounting for curvature or interactions in the **z**-vector.

In the second step, the adjusted weights $\tilde{w}_k$ are further modified so that survey-weighted estimates agree with known population totals. More specifically, we assume that a vector of calibration variables $\mathbf{x}^*$ is available for $k \in s_r$ and that the vector of population totals $\mathbf{t}_{\mathbf{x}^*} = \sum_{k\in U} \mathbf{x}_k^*$ is known. The $\mathbf{x}^*$-vector may contain one or more z-variables that were used in (8). The final weighting system is given by $\{w_k; k \in s_r\}$, where

$$w_k = \tilde{w}_k F\left(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k^*\right) \tag{10}$$

and $\hat{\boldsymbol{\lambda}}^\top$ is a vector of estimated coefficients. The final weights w_k satisfy the calibration constraints

$$\sum_{k\in s_r} w_k \mathbf{x}_k^* = \mathbf{t}_{\mathbf{x}}*. \tag{11}$$

The weight w_k in (10) is the product of the adjusted weight $\tilde{w}_k$ and the calibration adjustment factor $F\left(\hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_k^*\right)$.

For example, the linear method (4) leads to

$$w_k = \tilde{w}_k\left(1 + \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k^*\right),$$

whereas the exponential method leads to

$$w_k = \tilde{w}_k \exp\left(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k^*\right).$$

Alternative weighting methods are discussed in Kott and Liao (2012). Applying the final weighting system, $\{w_k;\, k \in s_r\}$, to a characteristic of interest y leads to the two-step calibration estimator

$$\hat{t}_{C,2} = \sum_{k\in s_r} w_k y_k = \sum_{k\in s_r} d_k \hat{\phi}_k^{-1} F\left(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k^*\right) y_k. \tag{12}$$

We make the following remarks: (i) if the nonresponse model (8) is correctly specified (and so the estimator (9) is asymptotically unbiased for t_y for every characteristic of interest), the two-step calibration estimator $\hat{t}_{C,2}$ is asymptotically unbiased for t_y regardless of the characteristic y being estimated. (ii) If the $\mathbf{x}^*$-vector is linearly related to y, then $\hat{t}_{C,2}$ is expected to be more efficient than $\hat{t}_{PSA}$. (iii) As for the complete data case, $\hat{\boldsymbol{\lambda}} \to 0$ in probability as $n \to \infty$ and $N \to \infty$ if the nonresponse model (8) is correctly specified. (iv) In the two-step approach, the calibration function is chosen using the same criteria as those encountered in the complete-data case. Most often, the distribution of the calibration adjustment factors $F\left(\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k^*\right)$ drives the choice of the function $F(\cdot)$.

4. The One-Step Approach: Nonresponse Calibration Weighting

Following Särndal and Lundström (2005), we distinguish between two levels of auxiliary information:

(1) U-level: a vector of auxiliary variables $\mathbf{x}_k^*$ is minimally available for $k \in s_r$ and the vector of population totals $\mathbf{t}_{\mathbf{x}^*} = \sum_{k \in U} \mathbf{x}_k^*$ is known.
(2) s-level: a vector of auxiliary variables $\mathbf{x}_k^o$ is available for $k \in s$ but the vector of population totals, $\sum_{k \in U} \mathbf{x}_k^o$, is unknown. Instead, the vector of complete-data estimators, $\hat{\mathbf{t}}_{\mathbf{x}^o} = \sum_{k \in s} d_k \mathbf{x}_k^o$, is available.

We define the stacked vector of auxiliary variables for unit k as $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$ and the corresponding vector of totals $\mathbf{t}_{\mathbf{x}} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}^*} \\ \hat{\mathbf{t}}_{\mathbf{x}^o} \end{pmatrix}$. The x^o-variables are believed to be associated with nonresponse and, possibly, with some characteristics of interest. Their role is similar to that of the z-variables in the two-step approach: contribute to reducing the nonresponse bias.

The final weighting system is $\{w_k;\ k \in s_r\}$, where

$$w_k = d_k F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right), \tag{13}$$

and $\hat{\boldsymbol{\lambda}}_r$ is determined so that the calibration constraints

$$\sum_{k \in s_r} w_k \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}$$

are satisfied. The final weight w_k in (13) is the product of the design weight d_k and the nonresponse/calibration adjustment factor $F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right)$. Applying the final weighting system, $\{w_k;\ k \in s_r\}$, to a characteristic of interest y leads to the one-step calibration estimator

$$\hat{t}_{C,1} = \sum_{k \in s_r} w_k y_k = \sum_{k \in s_r} d_k F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right) y_k. \tag{14}$$

Note that, unlike the two-step approach, the vector of estimated coefficient $\hat{\boldsymbol{\lambda}}_r$ does not converge towards 0 as $n \to \infty$ and $N \to \infty$. This is due to the fact that $F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right)$ is essentially an estimate of ϕ_k^{-1}.

We now compare the one-step and the two-step approaches. To that end, note that it is sufficient to compare the PSA estimator (which is the estimator resulting from the first step in the two-step approach) and a calibration estimator based on the x^o-variables only. The second step in the two-step approach or the use of the $\mathbf{x}^*$-variables in the one-step approach strive to make survey estimates and known population totals agree, which is not the focus here. Below, we argue that the one-step based on the x^o-variables imposes a parametric model for the relationship between the response propensity and the vector of auxiliary variables, which makes the resulting estimator vulnerable to a misspecification of the calibration function.

Recall that $\hat{t}_{PSA}$ is asymptotically unbiased for t_y regardless of the characteristic y being estimated, provided that the nonresponse model (8) is correctly specified. Therefore, for $\hat{t}_{C,1}$ in (14) to be asymptotically unbiased for t_y regardless of the characteristic y being estimated, we require

$$F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right) = \hat{\phi}_k^{-1}.$$

The previous expression suggests that the adjustment factor $F\left(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k\right)$ can be viewed as an implicit estimate of ϕ_k^{-1}.

Next, we examine the bias of $\hat{t}_{C,1}$, where the bias is defined as $Bias(\hat{t}_{C,1}) = E_pE_q(\hat{t}_{C,1}|s) - t_y$, and the subscripts p and q refer to the sampling design and the nonresponse mechanism respectively. Using a first-order Taylor expansion and ignoring the higher-order terms, the bias of $\hat{t}_{C,1}$ can be approximated by

$$Bias(\hat{t}_{C,1}) \approx -\sum_{k\in U}(1-\phi_k F_k)\left(y_{\mathrm{k}} - \mathbf{x}_k^\top \boldsymbol{B}_{\phi f}\right), \tag{15}$$

where

$$\boldsymbol{B}_{\phi f} = \left(\sum_{k\in U}\phi_k f_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1}\sum_{k\in U}\phi_k f_k \mathbf{x}_k y_k$$

with $F_k \equiv F\left(\boldsymbol{\lambda}_N^\top \mathbf{x}_k\right), f_k \equiv F'\left(\boldsymbol{\lambda}_N^\top \mathbf{x}_k\right)$ and $\boldsymbol{\lambda}_N$ denotes the probability limit of $\hat{\boldsymbol{\lambda}}_r$.

In the case of linear weighting (4), Expression (15) reduces to

$$Bias(\hat{t}_{C,1}) \approx -\sum_{k\in U}(1-\phi_k)\left(y_k - \mathbf{x}_k^\top \boldsymbol{B}_{\phi}\right), \tag{16}$$

where

$$\boldsymbol{B}_{\phi} = \left(\sum_{k\in U}\phi_k \mathbf{x}_k \mathbf{x}_{\mathrm{k}}^\top\right)^{-1}\sum_{k\in U}\phi_k \mathbf{x}_k y_k.$$

Expression (16) is identical to Expression (9.14) in Särndal and Lundström (2005). Note that the more general expression (15) does not appear in Särndal and Lundström (2005), where the focus is placed on linear weighting.

Expression (15) is interesting because it sheds some light on the conditions required for asymptotic unbiasedness:

(1) On the one hand, the asymptotic bias (15) vanishes if the finite population covariance between the residuals $e_k = \left(y_k - \mathbf{x}_k^\top \boldsymbol{B}_{\phi f}\right)$ and $\delta_k = \phi_k F_k - 1$ is equal to zero.

This condition is satisfied if

$$y_k = \mathbf{x}_k^\top \beta + \boldsymbol{\epsilon}_k \tag{17}$$

with

$$\mathrm{E}(\boldsymbol{\epsilon}_k|\mathbf{x}_k) = 0 \tag{18}$$

and if the response probability ϕ_k is not related to y_k after conditioning on $\mathbf{x}_k$. The latter condition is essentially the customary MAR assumption (Rubin 1976).

In multipurpose surveys, it is unrealistic to presume that Model (17) holds for every characteristic of interest y, in which case some estimates may suffer from bias. In fact, in household and social surveys, most characteristics of interest are categorical, in which case (17) is generally not appropriate.

(2) On the other hand, the asymptotic bias of $\hat{t}_{C,1}$ is equal to zero if

$$F_k = \phi_k^{-1}. \tag{19}$$

Hence, selecting a calibration function $F(\cdot)$ such that (19) is satisfied ensures that the one-step calibration estimator is asymptotically unbiased regardless of the characteristic of interest y being estimated, even if (17) and (18) do not hold. For linear weighting, it follows from (19) that $\hat{t}_{C,1}$ is asymptotically unbiased for t_y for every y if

$$\phi_k^{-1} = 1 + \boldsymbol{\lambda}^\top \mathbf{x}_k \quad \text{for all } k \in U, \tag{20}$$

for a vector of unknown constants $\boldsymbol{\lambda}$ (see Särndal and Lundström 2005, ch. 9). For exponential weighting, we require

$$\phi_k^{-1} = \exp\left(\boldsymbol{\lambda}^\top \mathbf{x}_k\right) \quad \text{for all } k \in U; \tag{21}$$

see also Kott and Liao (2012) for a discussion of alternative weighting methods. In other words, both the linear and exponential methods correspond to specific parametric nonresponse models, which suggests that selecting either one is somehow equivalent to (implicitly) selecting a nonresponse model. This begs the following question: how is $\hat{t}_{C,1}$ affected if (20) (respectively (21)) is not an appropriate description of the relationship linking the $\mathbf{x}$-vector and the ϕ_k's, that is, if the calibration function is misspecified? This aspect is investigated in Section 5.

A key aspect here is to realize that each calibration function corresponds to a specific parametric nonresponse model. By choosing a given calibration function, one is effectively making a strong statement about the underlying nonresponse mechanism. Therefore, in order to avoid an incorrect functional form, a complete modeling exercise is needed to validate the form of the function linking the response propensity ϕ_k to the vector of auxiliary variable $\mathbf{x}_k$. Failing to do so may result in biased estimators. Furthermore, there may be no calibration that corresponds to the inverse of the estimated response probabilities. For instance, suppose that the relationship between the response probability and a single auxiliary variable x is described by a nonmonotonic function. In this case, it may be difficult to find a calibration function that provides an adequate description of the relationship between the inverse of the response propensity and the $\mathbf{x}$-vector.

5. Simulation Study

We conducted a simulation study to illustrate the importance of carefully selecting a calibration function $F(\cdot)$ in the context of a one-step approach. We generated a population of size $N = 1{,}000$, which consisted of an auxiliary variable x and four variables of interest y_1, y_2, y_3 and y_4. The x-values were first generated from a uniform distribution (0, 80). The y_1-values were generated according to the linear model

$$y_{k1} = 1{,}000 + 10x_k + \varepsilon_{k1},$$

where the errors ε_{k1} were generated from a normal distribution with mean 0 and variance 300. The y_2-values were generated according to the exponential model

$$y_{k2} = exp(-0.1 + 0.1x_k) + \varepsilon_{k2},$$

where the errors ε_{k2} were generated from a normal distribution with mean 0 and variance 300. The y_3-values were generated according to the logistic model

$$y_{k3} \sim B(1, p_k),$$

where $p_k = [\exp\{-0.5\ \ (x_k - 55)\} + 1]^{-1}$. The y_4-values were generated according to the quadratic model

$$y_{k4} = 1{,}300 - (x_k - 40)^2 + \varepsilon_{k4},$$

where the errors ε_{k4} were generated from a normal distribution with mean 0 and standard deviation 300. The relationships between y_j and x are displayed in Figure 1, $j = 1, \ldots, 4$.

In order to focus on the nonresponse error, we considered the census case; that is, $n = N = 1{,}000$ and $d_k = 1$ for all k. In each population, units were assigned a response probability ϕ_k according to a given nonresponse mechanism. We simulated nonresponse according to four nonresponse mechanisms, all presented in Table 1; see also Figure 2. For each mechanism, the parameters were set so that the overall response rate was

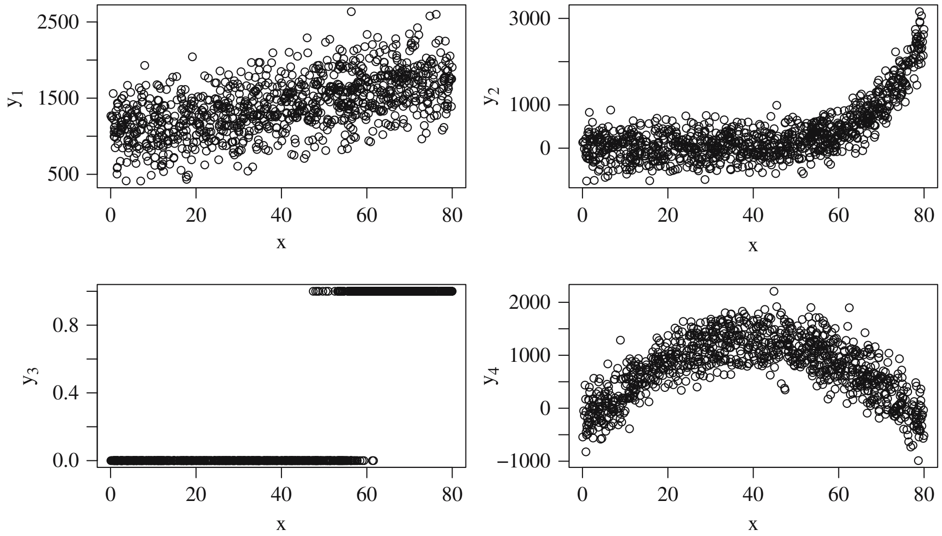


Fig. 1. Relationships between the characteristics of interest and x

Table 1. Nonresponse mechanisms used for generating nonresponse

Nonresponse mechanism	Name	ϕ_k
1	Inverse linear	$(1.2 + 0.024\ x_k)^{-1}$
2	Exponential	$\exp(-0.2 - 0.014x_k)$
3	Logistic type	$0.2 + 0.6\{1 + \exp(-5 + x_k/8)\}^{-1}$
4	Quadratic	$0.7 + 0.45\ (x_k/40 - 1)^2 + 0.0025\ x_k$

approximately equal to 50%. The response indicators R_k for $k \in U$ were generated independently from a Bernoulli distribution with parameter ϕ_k, resulting in a population of respondents U_r of size N_r. The nonresponse process was repeated $M = 5{,}000$ times, leading to $M = 5{,}000$ sets of respondents for each nonresponse mechanism. From Figure 1 and Figure 2, we note that both the response propensity and the characteristics of interest $y_1 - y_4$ are highly related to x in all the scenarios.

We were interested in estimating the populations totals $t_{y_j}, j = 1, 2, 3, 4$. For each total, we computed three estimators: (i) The unadjusted estimator $\hat{t}_{un} = N\bar{y}_r$ where $\bar{y}_r = \sum_{k\in U_r} y_k/N_r$; (ii) The one-step calibration estimator $\hat{t}_{C,1}$ given by (14) based on different calibration functions: linear, exponential and logit, given by (4), (5) and (7), respectively, using $\mathbf{x}_k = (1, x_k)$ as the auxiliary vector. In other words, the estimator $\hat{t}_{C,1}$ was calibrated on the population size N as well as the population total of x-values, t_x; (iii) the Propensity-Score Adjusted estimator $\hat{t}_{PSA}$, where the response propensities were estimated using the score method described in Section 3. To that end, preliminary response probabilities $\tilde{\phi}_k$ were first obtained using a logistic regression model with $(1, x_k)^\top$ as the vector of predictors. Then, the sample was partitioned into 20 weighting classes according the $\tilde{\phi}_k$'s and the response propensity of a unit in a given class was estimated using the response rate observed within the same class. Although five imputation classes are often sufficient for an effective bias reduction (Eltinge and Yansaneh 1997; Rosenbaum and

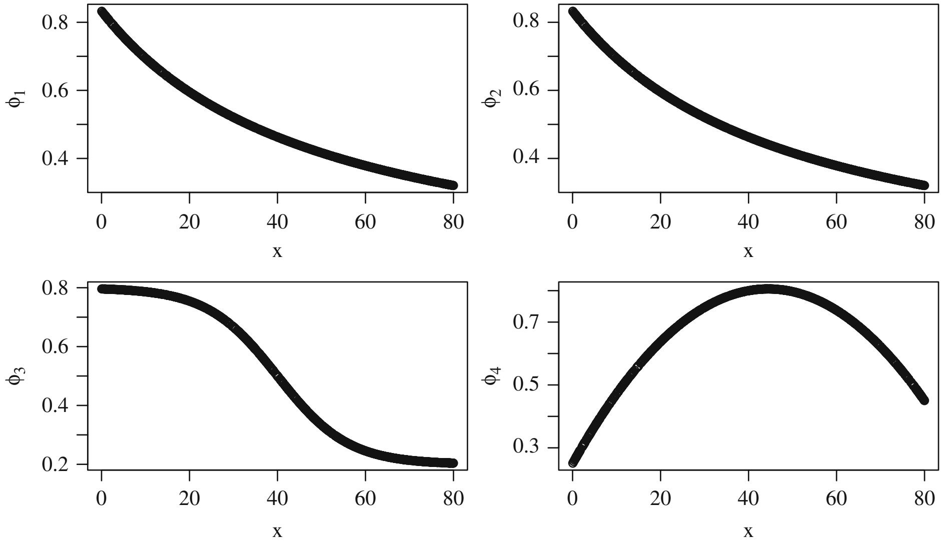


Fig. 2. Relationships between the response probability and x

Rubin 1983), it may not be appropriate when the relationship between the characteristic of interest and the auxiliary variable is highly nonlinear or contains a quadratic terms as it is the case for y_1 and y_4, respectively (see Haziza and Beaumont 2007). This is why we used 20 imputation classes.

As we argued in Section 4, in order to show that one-step calibration is vulnerable to the misspecification of the calibration function, it is sufficient to compare $\hat{t}_{PSA}$ and $\hat{t}_{C,1}$ based on the x^o-variables only. In other words, there is no need to perform the second step in the two-step approach or to use x*-type variables in the one-step approach.

As a measure of bias of an estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo percent relative bias (RB)

$$RB_{MC}(\hat{\theta}) = \frac{100}{M}\sum_{m=1}^{M}\frac{(\hat{\theta}_{(m)} - \theta)}{\theta},$$

where $\hat{\theta}_{(m)}$ denotes the estimator $\hat{\theta}$ in the m-th repetition, $m = 1, \ldots, M$. We also computed the percent relative root mean square error (RRMSE) of $\hat{\theta}$:

$$RRMSE_{MC}(\hat{\theta}) = 100 \times \frac{\left\{M^{-1}\sum_{m=1}^{M}(\hat{\theta}_{(m)} - \theta)^2\right\}^{1/2}}{\theta}.$$

The results are shown in Tables 2–5. As expected, the unadjusted estimator was biased in all the scenarios. This can be explained by the fact that the response probability was related to the characteristics of interest via the auxiliary variable x and that the unadjusted estimator did not account for x.

We now turn to the variable y_1, which was linearly related to the variable x. We see from Tables 2–5 that the resulting one-step calibration estimator $\hat{t}_{C,1}$ showed negligible bias regardless of the calibration method $F(\cdot)$ used. This is consistent with the Expressions (15)–(18). Furthermore, the choice of calibration method did not affect the efficiency of the estimator for a given nonresponse mechanism. For example, in Table 2

Table 2. Monte Carlo percent relative bias and percent relative root mean square error (in parenthesis) of several estimators under the inverse linear nonresponse mechanism: $\Phi_k = (1.2 + 0.024\ x_k)^{-1}$

		$\hat{t}_{yC,1}$ $F(u) =$			
	$\hat{t}_{un}$	$1+u$	$\exp(u)$	$\dfrac{L(M-1)+M(1-L)\exp(Au)}{M-1+(1-L)\exp(Au)}$	$\hat{t}_{PSA}$
y_1 (linear)	−4.1 (4.2)	0.0 (0.7)	0.0 (0.7)	0.0 (0.7)	−0.0 (0.8)
y_2 (exponential)	−28.1 (28.7)	−0.1 (5.5)	2.8 (6.1)	3.3 (6.4)	−0.1 (3.0)
y_3 (logistic)	−27.5 (27.9)	−0.1 (3.4)	1.7 (3.6)	2.1 (3.8)	−0.1 (2.3)
y_4 (quadratic)	−4.8 (5.3)	0.1 (2.8)	−2.0 (3.3)	−2.4 (3.5)	−0.1 (1.4)

Table 3. Monte Carlo percent relative bias and percent relative root mean square error (in parenthesis) of several estimators under the exponential nonresponse mechanism: $\Phi_k = exp(-0.2 - 0.014x_k)$

		$\hat{t}_{yC,1}$ $F(u) =$			
	$\hat{t}_{un}$	$1+u$	$\exp(u)$	$\dfrac{L(M-1)+M(1-L)\exp(Au)}{M-1+(1-L)\exp(Au)}$	$\hat{t}_{PSA}$
y_1	-4.9	-0.0	0.0	0.0	-0.0
(linear)	(4.9)	(0.8)	(0.8)	(0.8)	(0.8)
y_2	-35.1	-4.0	-0.0	0.7	-0.1
(exponential)	(35.5)	(7.1)	(5.8)	(5.9)	(3.2)
y_3	-33.8	-2.5	0.0	0.6	-0.1
(logistic)	(34.1)	(4.3)	(3.3)	(3.3)	(2.3)
y_4	-3.6	2.9	0.0	-0.6	-0.0
(quadratic)	(4.3)	(4.2)	(2.7)	(2.8)	(1.6)

(which corresponds to the inverse linear nonresponse mechanism), the RRMSE of $\hat{t}_{C,1}$ was equal to 0.7 for all the calibration methods. Finally, the PSA estimator showed virtually no bias in all the scenarios corresponding to the y_1-variable and showed the same efficiency as that of $\hat{t}_{C,1}$, except in Table 2, where we note a slight loss of efficiency.

For the variables y_2-y_4 that were not linearly related to the x-variable, we note that the resulting one-step calibration estimator was generally biased, except when the calibration method $F(\cdot)$ was appropriate; see Expression (19). For example, in Table 2 (which corresponds to the inverse linear nonresponse mechanism), the one-step calibration estimator $\hat{t}_{C,1}$ showed no bias for the three variables under the linear calibration method $F(u) = 1 + u$. These results are consistent with (20). On the other hand, the other calibration methods (exponential and logit) led to some bias with an absolute RB ranging

Table 4. Monte Carlo percent relative bias and percent relative root mean square error (in parenthesis) of several estimators under the logistic nonresponse mechanism: $\Phi_k = 0.2 + 0.6\{1 + exp(-5 + x_k/8)\}^{-1}$

		$\hat{t}_{yC,1}$ $F(u) =$			
	$\hat{t}_{un}$	$1+u$	$\exp(u)$	$\dfrac{L(M-1)+M(1-L)\exp(Au)}{M-1+(1-L)\exp(Au)}$	$\hat{t}_{PSA}$
y_1	-7.3	-0.3	-0.2	-0.2	-0.1
(linear)	(7.3)	(0.9)	(0.9)	(0.9)	(0.9)
y_2	-51.5	-10.0	-0.4	0.9	-0.2
(exponential)	(51.7)	(12.3)	(7.0)	(7.1)	(3.7)
y_3	-53.4	-12.1	-5.6	-4.5	-0.3
(logistic)	(53.5)	(12.9)	(6.7)	(5.8)	(3.0)
y_4	-1.0	11.7	4.3	3.1	-0.0
(quadratic)	(2.3)	(12.3)	(5.3)	(4.3)	(1.8)

Table 5. Monte Carlo percent relative bias and percent relative root mean square error (in parenthesis) of several estimators under the quadratic nonresponse mechanism: $\Phi_k = 0.7 + 0.45\ (x_k/40 - 1)^2 + 0.0025\ x_k$

			$\hat{t}_{yC,1}$ $F(u) =$		
	$\hat{t}_{un}$	$1 + u$	$\exp(u)$	$\dfrac{L(M-1) + M(1-L)\exp(Au)}{M - 1 + (1-L)\exp(Au)}$	$\hat{t}_{PSA}$
y_1 (linear)	1.3 (1.4)	− 0.2 (0.5)	− 0.2 (0.5)	− 0.2 (0.5)	0.0 (0.5)
y_2 (exponential)	− 8.3 (9.4)	− 19.7 (19.9)	− 19.2 (19.5)	− 19.0 (19.3)	− 0.4 (2.3)
y_3 (logistic)	− 0.5 (3.2)	− 11.8 (12.0)	− 11.5 (11.7)	− 11.4 (11.6)	− 0.1 (1.0)
y_4 (quadratic)	13.1 (13.2)	13.7 (13.8)	13.4 (13.5)	13.2 (13.4)	0.2 (1.1)

from 1.7% to 3.3%. Similarly, in Table 3 (which corresponds to the exponential nonresponse mechanism), the one-step calibration estimator $\hat{t}_{C,1}$ showed no bias for the three variables under the exponential calibration method $F(u) = \exp(u)$. These results are consistent with (21). On the other hand, the other calibration methods (linear and logit) led to some bias with an absolute RB ranging from 0.6% to 4.0%.

In Tables 4 and 5, we note that the one-step calibration estimator showed some bias in all the scenarios, which can be explained by the fact that none of the calibration methods (linear, exponential or logit) provided an adequate description of the relationship between the inverse of the response probability and the x-variable. For example, in Table 5, all the calibration methods led to substantial bias with an absolute RB ranging from 11.4% to 19.7%. It is worth noting that the one-step calibration estimator was significantly more biased than the unadjusted estimator for the variables y_2 and y_3, which illustrates that a poor choice of $F(\cdot)$ may result in significant biases, which can be larger than that of the unadjusted estimator. Finally, the PSA estimator showed negligible biases in all the scenarios corresponding to y_2-y_4. Moreover, its RRMSE was considerably smaller than that of the one-step calibration estimator for these variables. These results suggest that the score method, which is nonparametric in nature, is robust to the misspecification of the form of the function $m(\cdot)$ in (8).

The results presented here suggest that a high association between the characteristic variable and the auxiliary variables is not necessarily enough for the one-step calibration method to yield good results, as in the cases of the variables y_2 and y_3. Also, as shown in Table 5, the fact that various calibration functions yield about the same estimate is not necessarily a sign that any of the choices will work well.

6. Discussion

In this article, we have discussed two weighting approaches in the presence of unit nonresponse: the one-step approach and the two-step approach, the latter being the customary approach to weighting in statistical agencies. Although it is more complex to implement than the one-step approach as two distinct weighting procedures must be applied, the two-step approach offers several advantages: first, it makes it possible to assess the impact of nonresponse adjustment and calibration adjustment on the distribution of the weights separately. Furthermore, when multiple characteristics are collected, survey statisticians prefer modeling the response probability to the survey as it does not require a different model for each characteristic of interest. In this case, complete reliance is placed on the nonresponse model in order to achieve an efficient bias reduction for every y. In statistical agencies, the response propensities are typically estimated through nonparametric methods such as weighting classes based on estimated response probabilities or regression trees, as both types of methods provide protection against misspecification of the functional and account for curvature and interactions. This is especially important when the auxiliary variables are continuous and their association with the response probability is not monotonic.

In contrast, the single-step calibration approach is simple to implement as the whole weighting process is performed in a single step. Furthermore, it does not make explicit use of estimated response probabilities, unlike the two-step approach. However, as we have illustrated empirically, the choice of the calibration function $F(\cdot)$ is generally important. In the simulation study conducted in Section 5, where we considered the case of a quantitative variable x, the results suggested that the one-step calibration estimator suffered from significant bias if the calibration function is inappropriate. Would the results be similar if the calibration results were categorical? We revisit the case of two categorical variables x_1 and x_2 described in Section 2. Let $N_{j_1 j_2}$ be the individual cell counts available at the sample level (Info-s). Matching the individual cell counts results in a poststratified-type estimator, in which case the choice of the calibration function is unimportant as different $F(\cdot)$ would result in the same estimator. In other words, as long as the variables x_1 and x_2 are related to nonresponse, the one-step calibration estimator should exhibit no bias. In fact, in this case, the latter is identical to the PSA estimator based on weighting classes obtained by cross classifying x_1 and x_2. On the other hand, if calibration is performed to match the margins $N_{j_1 \bullet}$ and $N_{\bullet j_2}$ available at the sample level, choosing the appropriate calibration function becomes an issue once again, as different $F(\cdot)$ would lead to different one-step calibration estimators. In their Remark 10.1, Särndal and Lundström (2005) suggest that categorizing the x-variables when the latter are quantitative may bring some robustness. For a poststratification-type situation, we agree with this recommendation. However, when calibration is performed on margins only, the extent to which the one-step calibration estimators would be robust to the misspecification of the calibration function $F(\cdot)$ is not so clear-cut.

Although the PSA estimator based on the score method performed well in all the scenarios presented in Section 5, we are not suggesting that it would perform well in any type of situations. If a causal relationship exists between one or more characteristics of interest and the response propensity, some residual nonresponse bias will remain. Furthermore, we have considered the case of a single quantitative variable x. Additional studies are needed to investigate how the score method would perform in the presence of multiple quantitative variables with, possibly, quadratic or cubic terms. The results simply suggest that nonparametric methods are attractive from a practical point of view as they bring some robustness if the nonresponse model is not correctly specified. This is not true, in general, for the one-step approach that imposes an implicit parametric model.

7. References

Brick, M.J. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29: 329–353. Doi: http://dx.doi.org/10.2478/jos-2013-0026.

Da Silva, D.N. and J.D. Opsomer. 2006. "A Kernel Smoothing Method of Adjusting for Unit Non-Response in Sample Surveys." *The Canadian Journal of Statistics* 34: 563–579.

Da Silva, D.N. and J.D. Opsomer. 2009. "Nonparametric Propensity Weighting for Survey Nonresponse through Local Polynomial Regression." *Survey Methodology* 35: 165–176.

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.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020. Doi: http://dx.doi.org/10.1080/01621459.1993.10476369.

Ekholm, A. and S. Laaksonen. 1991. "Weighting via Response Modeling in the Finnish Household Budget Survey." *Journal of Official Statistics* 7: 325–337.

Eltinge, J.L. and I.S. Yansaneh. 1997. "Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey." *Survey Methodology* 23: 33–40.

Giommi, A. 1987. "Nonparametric Methods for Estimating Individual Response Probabilities." *Survey Methodology* 13: 127–134.

Haziza, D. and J.-F. Beaumont. 2007. "On the Construction of Imputation Classes in Surveys." *International Statistical Review* 75: 25–43. Doi: http://dx.doi.org/10.1111/j.1751-5823.2006.00002.x.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19: 81–97.

Kim, J.K. and J.J. Kim. 2007. "Nonresponse Weighting Adjustment Using Estimated Response Probability." *The Canadian Journal of Statistics* 35: 501–514. Doi: http://dx.doi.org/10.1002/cjs.5550350403.

Kott, P. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Undercoverage." *Survey Methodology* 32: 133–142.

Kott, P.S. and D. Liao. 2012. "Providing Double Protection for Unit Nonresponse With a Nonlinear Calibration-Weighting Routine." *Survey Research Methods* 6: 105–111.

Lee, S. 2006. "Propensity Score Adjustments as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22: 329–349.

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

Little, R.J.A. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Phipps, P. and D. Toth. 2012. "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model With Linked Administrative Data." *Annals of Applied Statistics* 6: 772–794. Doi: http://dx.doi.org/10.1214/11-AOAS521.

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

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–590. Doi: http://dx.doi.org/10.1093/biomet/63.3.581.

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. 2005. *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 147–164, http://dx.doi.org/10.1515/JOS-2016-0007

A Note on the Effect of Data Clustering on the Multiple-Imputation Variance Estimator: A Theoretical Addendum to the Lewis et al. article in JOS 2014

Yulei He[1], *Iris Shimizu*[1], *Susan Schappert*[1], *Jianmin Xu*[1], *Vladislav Beresovsky*[1], *Diba Khan*[1], *Roberto Valverde*[1], *and Nathaniel Schenker*[1]

Multiple imputation is a popular approach to handling missing data. Although it was originally motivated by survey nonresponse problems, it has been readily applied to other data settings. However, its general behavior still remains unclear when applied to survey data with complex sample designs, including clustering. Recently, Lewis et al. (2014) compared single- and multiple-imputation analyses for certain incomplete variables in the 2008 National Ambulatory Medicare Care Survey, which has a nationally representative, multistage, and clustered sampling design. Their study results suggested that the increase of the variance estimate due to multiple imputation compared with single imputation largely disappears for estimates with large design effects. We complement their empirical research by providing some theoretical reasoning. We consider data sampled from an equally weighted, single-stage cluster design and characterize the process using a balanced, one-way normal random-effects model. Assuming that the missingness is completely at random, we derive analytic expressions for the within- and between-multiple-imputation variance estimators for the mean estimator, and thus conveniently reveal the impact of design effects on these variance estimators. We propose approximations for the fraction of missing information in clustered samples, extending previous results for simple random samples. We discuss some generalizations of this research and its practical implications for data release by statistical agencies.

Key words: Bayesian; complex survey design; data release; exploratory data analysis; fraction of missing information; missing data.

1. Introduction

Data collected for scientific research often contain missing values. For example, the National Ambulatory Medical Care Survey (NAMCS) has been conducted by the U.S. Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) since 1973. The survey aims to provide nationally representative data on office-based physician care. The ultimate sample unit is a doctor-patient encounter, drawn systematically from the terminus of a multistage, clustered sample design. However, NAMCS has considerable item nonresponse for race, one of the key demographics used in various analyses. These missing data, if inadequately accounted for, might lead to invalid inferences and misleading policy implications.

[1] National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, 20782, U.S.A. Emails: wdq7@cdc.gov, ims1@cdc.gov, sds0@cdc.gov, ewl4@cdc.gov, hvy4@cdc.gov, ild1@cdc.gov, rcv4@cdc.gov, and nhs1@cdc.gov

Acknowledgments: The authors would like to thank Alan Dorfman, Donald Malec, and Jennifer Madans for their valuable input.

Multiple imputation (MI) (Rubin 1987) is a popular approach to handling missing data problems. In general, MI involves replacing each missing datum with several (D) sets of plausible values drawn from a specified imputation model, resulting in several completed datasets (i.e., data with missing values filled in by imputations). Each completed dataset is analyzed separately by a standard complete-data method. The resulting inferences, including point estimates, covariance matrices, and p-values, can then be combined to formally incorporate imputation uncertainty using the formulae given in Chapter 3 of Rubin (1987) and refined in Chapter 10 of Little and Rubin (2002). See also Subsection 2.3 for more specifics. The implementation of MI in several major statistical packages, including SAS (www.sas.com), R (www.r-project.org), and STATA (www.stata.com), has made this missing data strategy increasingly popular among practitioners (Harel and Zhou 2007).

MI was originally proposed as a Bayesian, model-based approach to survey nonresponse issues (Rubin 1978). However, it has been widely applied to data of various types such as surveys, clinical trials, and observational studies. Despite its popularity, limited research has been conducted to assess the general behavior of MI for survey data with complex sample designs such as stratification and clustering. Rubin (1987, chap. 4) provided some general, asymptotic arguments for the appropriateness of MI for survey data. Rubin and Schenker (1986) used data from simple random samples (SRS) as illustrations. Meng (1995) raised the issue of "uncongeniality" for MI inferences, which occurs when imputation models might be incompatible with complete-data analysis procedures. See also Kim and Shao (2014, chap. 4) for further discussion of this topic. Reiter et al. (2006) demonstrated that bias can arise when complex survey-design features are not accounted for in the imputation models.

Recently, Lewis et al. (2014) applied MI to the race variable (around 30% missing) in the 2008 NAMCS and estimated race proportions at national and domain levels. They compared the variance estimates from MI and single imputation (SI), and the study results suggested that the variance increase due to MI decreases as the design effects of the estimated proportions increase. That is, estimates with larger design effects are associated with smaller increases in estimated variance after MI, despite having similar rates of missingness. Similar patterns can also be identified in simulation studies conducted by Reiter et al. (2006). It is generally expected that the variance increase due to MI is small when the rate of missingness is small. That is, there is little difference for the variance estimates between MI and SI when there is little missing data. However, the additional role played by the design effect in MI variance estimation is unclear. This phenomenon was termed "surprising" in Lewis et al. (2014), yet no convincing theoretical justification was provided. Fully understanding the rationale behind this phenomenon is important, given the increasing number of applications of MI to complex survey data (e.g., Schenker et al. 2006). This issue is also related to the emerging topics of research on conducting MI for other types of data (e.g., clinical trials) with clustered (multilevel) structure (see van Buuren 2012, sec. 3.8 and references therein).

In this article, we aim to provide some theoretical explanation as a complement to the empirical study in Lewis et al. (2014). We elucidate the effect of data clustering on MI variance estimation by deriving algebraic expressions and discuss its practical implications. The remainder of the article is organized as follows. In Section 2, we introduce a classic one-way normal random-effects model for balanced data to characterize the data sampled from a clustered design. This model-based setup is convenient for studying the properties of MI inference. In Section 3, we derive formulae for the between- and within-imputation variance components of MI analysis for the mean estimate under this model. The variance increase due to MI is shown to decrease as clustering (design) effects increase. Approximations for the fraction of missing information are proposed. Finally, in Section 4, we propose topics for future research.

2. Method

2.1. *Complete-Data Model*

Complex survey designs (e.g., in NAMCS) often include multistage stratification and clustering. It is often difficult to characterize such a process using explicit models. For simplicity, we consider single-stage cluster sampling with clusters of equal sizes (Cochran 1977, chap. 9). That is, a simple random sample of m clusters, each containing n elements, is drawn from M clusters in the population. We further consider a model-based representation of this sample as follows:

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \epsilon_{ij}, \\ \alpha_i &\overset{i.i.d.}{\sim} N(0, \tau^2), \\ \epsilon_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2), \end{aligned} \tag{1}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n$, where y_{ij} is the random variable, μ is the (super) population mean, the α_is are between-cluster random effects, and the ϵ_{ij}s represent within-cluster measurement error, and i.i.d. means "independent and identically distributed".

Model (1) (a balanced, one-way normal random-effects model) and its variants are frequently used in the analysis of clustered surveys (Valliant et al. 2000, chap. 8). Here we use Model (1) as a basis to derive the corresponding MI variance estimators and relate them to the design effects (Kish 1965) used in survey sampling. Model (1) and its generalizations, the mixed-effects models, are also used in the emerging literature on conducting MI for clustered data not limited to surveys (e.g., see Andridge 2011 for clustered randomized trials and Schafer and Yucel 2002 for longitudinal data).

Under Model (1), $Cov(y_{ij}, y_{ij'}) = \tau^2$ for $j \neq j'$ and $j, j' \in (1, \ldots, n)$, and $Cov(y_{ij}, y_{i'k}) = 0$ for $i \neq i'$, $i, i' \in (1, \ldots, m)$ and $j, k \in (1, \ldots, n)$. With complete data, the typical unbiased estimator with minimum variance for μ is the overall sample mean $\hat{\mu}_{com} = \bar{y}_{\cdot\cdot,com} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}}{mn}$. The variance of the estimator is $Var(\hat{\mu}_{com}|\tau^2, \sigma^2) = \tau^2/m + \sigma^2/mn$. Its unbiased variance estimator is $\frac{\sum_{i=1}^{m}(\bar{y}_{i\cdot,com} - \bar{y}_{\cdot\cdot,com})^2}{m(m-1)}$, where $\bar{y}_{i\cdot,com} = \sum_{j=1}^{n} y_{ij}/n$ is the sample average at the cluster level. On the other hand, if we were to wrongly ignore the within-cluster correlation and assume that $y_{ij} \overset{i.i.d.}{\sim} N(\mu, \tau^2 + \sigma^2)$, then the variance for the overall mean would be $(\tau^2 + \sigma^2)/mn$ under the misspecified model. The ratio between the two variances is $\frac{\tau^2/m + \sigma^2/mn}{(\tau^2 + \sigma^2)/mn} = 1 + (n-1)\rho$, where $\rho = \tau^2/(\tau^2 + \sigma^2)$ is the intraclass correlation (coefficient). From the perspective of design-based inference, the factor $1 + (n-1)\rho$ is the design effect, showing how much the variance is changed by the use of cluster sampling instead of SRS. We let $deff_{com} = 1 + (n - 1)\rho$, where *deff* denotes "design effect" as in survey statistics literature (e.g., Cochran 1977, 242; Valliant et al. 2013, 5). This design effect can also be interpreted as a model-based mispecification effect (Skinner et al. 1989, chap. 2).

Note that Model (1) ignores other features in typical complex survey data such as stratification, unequal cluster sizes, as well as multistage sampling. However, the simple expression for the design effect is useful for illustrating its connection with MI variance estimation. The limitations of Model (1) are discussed in Section 4.

2.2. *Missing Data*

Suppose that missing data occur in the original sample. For ease of notation, we assume that within cluster i, the first r_i out of the n observations are observed. That is, y_{ij}s are observed for $i = 1, \ldots, m, j = 1, \ldots, r_i, r_i < n$ and missing otherwise. Following Rubin and Schenker (1986), we assume that the missingness is completely at random (MCAR) (Little and Rubin 2002) for this univariate missing data problem. This simplified assumption allows us to focus on the effect of clustering alone, excluding predictive covariates from Model (1). Under MCAR, $E(r_i) = r$ for $i = 1, \ldots, m$, also implying that the missingness is unrelated to the clustering factor. See Section 4 for discussion related to a more general assumption for the nonresponse mechanism such as missing at random (MAR).

For simplicity of derivation, we let $r_i = r$ for $i = 1, \ldots, r$. The rate of missingness is therefore $(n - r)/n$. Under Model (1), it is easy to verify that the grand mean of the observed data $\hat{\mu}_{obs} = \bar{y}_{\cdot\cdot,obs} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij}}{mr}$ is unbiased: $E(\hat{\mu}_{obs}) = \mu$. Its variance is $Var(\bar{y}_{\cdot\cdot,obs}|\tau^2, \sigma^2) = \tau^2/m + \sigma^2/mr$, and an unbiased variance estimator is $\frac{\sum_{i=1}^{m}(\bar{y}_{i\cdot,obs} - \bar{y}_{\cdot\cdot,obs})^2}{m(m-1)}$, where $\bar{y}_{i\cdot,obs} = \sum_{j=1}^{r} y_{ij}/r$ is the observed-sample mean at the cluster level. Therefore the design effect based on the observed data is $deff_{obs} = \frac{\tau^2/m + \sigma^2/mr}{(\tau^2 + \sigma^2)/mr} = 1 + (r - 1)\rho$.

2.3. *A Brief Review of MI*

Before we present more specifics, we briefly review the MI framework from a Bayesian model-based perspective. For an incomplete dataset $Y = \{Y_{obs}, Y_{mis}\}$, where Y_{obs} and Y_{mis} denote the observed and missing components of Y, respectively, we are interested in estimating a (scalar) population quantity Q. From the perspective of model-based inference, Q can often be treated as a superpopulation parameter in a posited model (e.g., μ in Model (1)). We further assume that the missingness is at random, which means that the probability of missingness is only related to fully observed variables or is some constant, the latter case being MCAR as a special case of MAR. According to Rubin (1987, chap. 3) and Little and Rubin (2002, sec. 10.2.1), the underlying theory behind MI analysis is

$$E(Q|Y_{obs}) = E_{Y_{mis}} E(Q|Y_{obs}, Y_{mis}), \tag{2}$$

$$Var(Q|Y_{obs}) = Var_{Y_{mis}} E(Q|Y_{obs}, Y_{mis}) + E_{Y_{mis}} Var(Q|Y_{obs}, Y_{mis}),$$

where Y_{mis} (the missing values for which imputations are created) are drawn from their posterior predictive distributions $P(Y_{mis}|Y_{obs})$.

In the imputation stage of MI, we draw Y_{mis} independently D times to create D completed datasets. Let $\hat{Q}$ denote the complete-data estimate for Q. In the analysis stage, the MI estimator for Q is $\hat{Q}_{MI} = \frac{\sum_{d=1}^{D} \hat{Q}^{(d)}_{Y_{obs}, Y^{(d)}_{mis}}}{D}$ (the average of $\hat{Q}$ evaluated using the completed datasets). Its variance is estimated by a weighted sum of the average within-imputation variance and the between-imputation variance. That is, $Var(\hat{Q}_{MI}) = W + \left(1 + \frac{1}{D}\right)B$, where $W = \frac{\sum_{d=1}^{D} Var(\hat{Q}^{(d)}_{Y_{obs}, Y^{(d)}_{mis}})}{D}$ the average within-imputation variance), and $B = \frac{\sum_{d=1}^{D} (\hat{Q}^{(d)}_{Y_{obs}, Y^{(d)}_{mis}} - \hat{Q}_{MI})^2}{D-1}$ (the between-imputation variance). The coefficient of B, that is, $1 + \frac{1}{D}$, approaches 1 as $D \rightarrow \infty$. Rubin (1987) argued that as $D \rightarrow \infty$, $\hat{Q}_{MI} \rightarrow E(Q|Y_{obs})$, $W \rightarrow E_{Y_{mis}} Var(Q|Y_{obs}, Y_{mis})$, and $B \rightarrow Var_{Y_{mis}} E(Q|Y_{obs}, Y_{mis})$.

The increase of variance due to the use of MI instead of SI (Lewis et al. 2014) can be alternatively quantified using the fraction of missing information (FMI) (Rubin 1987), a key element of MI analysis output. FMI is approximately the ratio of between-imputation variance to total variance; $FMI \approx B/(B + W)$, with the approximate equality approaching exact equality as $D \rightarrow \infty$, also termed as the population fraction of missing information (Rubin 1987, 86 and 114). It typically depends to some extent on the percent of missingness. It also depends on the analysis of interest and the extent to which the imputation model is predictive of the missing values. For example, for a univariate missing data problem with no covariate in the imputation model, the FMI for the mean estimator is approximately the rate of missingness (Rubin 1987, 114). However, if the imputation model includes other predictive covariates, the FMI will tend to be smaller than the item nonresponse rate, reflecting the gain in precision by using these covariates.

For brevity and clarity, we mainly consider the scenario with an infinite number of imputations ($D \rightarrow \infty$). We discuss relevant issues with a finite number of imputations in Section 4.

3. MI Variance Estimators under Model (1)

3.1. *The Effect of Design Effects*

We aim to relate the design effect to FMI in the scenario considered in Subsections 2.1 and 2.2. Let the imputed values from the dth imputation be $y_{ij}^{*(d)}$, $i = 1, \ldots, m$, $j = r + 1, \ldots, n$ and $d = 1, \ldots, D$. Then for each completed dataset $\{\{y_{ij}\}, \{y_{ij}^{*(d)}\}\}$, the completed-data estimator is $\hat{\mu}^{(d)}_{com} = \bar{y}^{(d)}_{\cdot\cdot,com} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij} + \sum_{i=1}^{m}\sum_{j=r+1}^{n} y_{ij}^{*(d)}}{mn}$. The MI estimator for μ is $\hat{\mu}_{MI} = \sum_{d=1}^{D} \hat{\mu}^{(d)}_{com}/D$. For the dth dataset, the within-imputation variance estimator is $\frac{\sum_{i=1}^{m}\left(\bar{y}^{(d)}_{i\cdot,com} - \bar{y}^{(d)}_{\cdot\cdot,com}\right)^2}{m(m-1)}$, where $\bar{y}^{(d)}_{i\cdot,com} = \frac{\sum_{j=1}^{r} y_{ij} + \sum_{j=r+1}^{n} y_{ij}^{*(d)}}{n}$. The between-imputation variance estimator is $\frac{\sum_{d=1}^{D}(\hat{\mu}^{(d)}_{com} - \hat{\mu}_{MI})^2}{D-1}$.

In the Appendix, we consider two MI scenarios, one in which τ^2 and σ^2 are known and the other in which they are unknown and require estimation that is embedded in the imputation. In both cases, it is shown that as $D \longrightarrow \infty$, $E(\hat{\mu}_{MI}) = \mu$, $Var(\hat{\mu}_{MI}) \longrightarrow \tau^2/m + \sigma^2/mr$, $E(W) \longrightarrow \tau^2/m + \sigma^2/mn$, and $E(B) \longrightarrow \sigma^2/mr - \sigma^2/mn$.

Under Model (1), the MI estimator is asymptotically equivalent to the complete-case estimator $\hat{\mu}_{obs}$ (Subsection 2.2). This is expected because of the MCAR mechanism and no predictive covariate is included in the imputation model. This is also consistent with the case of SRS (Rubin and Schenker 1986). In addition, the expected within-imputation variance $E(W)$ is asymptotically identical to $Var(\hat{\mu}_{com})$ as if data were not missing (Subsection 2.1). This makes intuitive sense because under a correctly specified model, the imputations are expected to retain the features of the unobserved data. Therefore the completed-data statistics shall preserve the mean and variance structure of the original, complete data. One might reasonably question the necessity of MI in this case. However, the explicit expressions for $E(B)$ and $E(W)$ shed some light on the effect of clustering on MI variance estimation.

Note that as $D \longrightarrow \infty$, $FMI \longrightarrow \frac{E(B)}{Var(\hat{\mu}_{MI})} = \frac{\sigma^2/mr - \sigma^2/mn}{\tau^2/m + \sigma^2/mr} = \frac{\frac{n-r}{nr}(1-\rho)}{\rho + \frac{1}{r}(1-\rho)}$. Plugging in $\rho = (deff_{obs} - 1)/(r-1) = (deff_{com} - 1)/(n-1)$ and we can shown that, in the limit,

$$FMI = \frac{n-r}{n}\frac{r - deff_{obs}}{(r-1)deff_{obs}} = \frac{n-r}{n}\frac{n - deff_{com}}{n - r + (r-1)deff_{com}}.$$

Let $n \longrightarrow \infty$ (so that $r \longrightarrow \infty$ for a fixed missingness rate). Then

$$FMI = \frac{n-r}{n}\frac{r - deff_{obs}}{(r-1)deff_{obs}} = \frac{n-r}{n}\left[\frac{1}{\left(1-\frac{1}{r}\right)deff_{obs}} - \frac{1}{r-1}\right] \approx \frac{n-r}{n}\frac{1-\frac{deff_{obs}}{r}}{deff_{obs}},$$

and similarly,

$$FMI \approx \frac{n-r}{n}\frac{1-\frac{deff_{com}}{n}}{\frac{n-r}{n}+\frac{r}{n}deff_{com}}.$$

However, practical surveys might have more complicated designs than the one-stage cluster design that we consider. Thus it might be difficult to pinpoint n and r in those contexts. To make the derived relationship widely useful, we aim to obtain expressions that only involve the rate of missing data $\left(P_{mis} = \frac{n-r}{n}\right)$ and design-effect estimates, both of which are readily available for general surveys. Therefore we consider the following simplifications:

$$FMI \approx \frac{P_{mis}}{deff_{obs}}, \tag{3}$$

and

$$FMI \approx \frac{P_{mis}}{(1-P_{mis})deff_{com} + P_{mis}}, \tag{4}$$

where P_{mis} quantifies the rate of missingness.

Note that Approximations (3) and (4) can be viewed as further approximations if $deff_{obs} \ll r$ and $deff_{com} \ll n$ (i.e., the design effects are much less than the cluster size). Otherwise we can treat them as upper bounds which are simple to calculate. We use Approximations (3) and (4) in the numerical illustrations (Subsection 3.2) and discuss their practical use.

Moreover, the approximations are derived for clustered data, including SRS as a special case. In the latter scenario, $\rho = 0 \Rightarrow deff_{obs} = deff_{com} = 1$, and thus $FMI \approx P_{mis}$, matching the results stated in Rubin (1987, 114). For data with a fixed missingness rate, Approximations (3) and (4) imply that FMI decreases as $deff_{obs}$ or $deff_{com}$ increases, explaining the phenomenon identified in Lewis et al. (2014).

In the example considered in this section, the variance of the infinite$-D$ MI estimator is a sum of the between- and within-cluster variance, that is, τ^2/m and σ^2/mr. When the intraclass correlation (or design effect) increases, the between-cluster variance dominates the within-cluster variance. Correspondingly in MI, the imputations from each cluster can be viewed as draws around the corresponding cluster average (i.e., $y^*_{ij} \sim N(\bar{y}_{i\cdot,obs}, (1+1/r)\sigma^2)$; see the Appendix). Thus the associated uncertainty, which is reflected by the between-imputation variance B, is only of the magnitude of the within-cluster variance σ^2, implying that the between-imputation variance contributes little to the total variance.

Although Approximations (3) and (4) are derived under the same Model (1), their uses in more general scenarios might yield different results. Practically, Approximation (3) can be calculated using the incomplete cases, while Approximation (4) can only be calculated using imputed data (because we do not have complete data), assuming that the imputation model adequately captures the complete-data structure and relationships. It is also plausible that the approximations do not always agree when both the design and missingness mechanisms of the survey data are more complicated than what we assume in Model (1).

3.2. *Numerical Illustrations*

Subsection 3.1 presents some theoretical derivations under a simple one-stage clustering design. As a follow-up study to Lewis et al. (2014), we assess the practical applicability of our theoretical results (i.e., Approximations (3) and (4)) by comparing them with real-study results of Lewis et al. (2014). Since the NAMCS data have a more complicated sample design and nonresponse mechanism, we expect to see both agreements and disagreements.

Lewis et al. (2014) estimated the ratio of the standard errors between MI and SI and used it as a metric to summarize the main findings. This ratio is a monotonic transformation of FMI as $SE(\hat{\mu}_{MI})/SE(\hat{\mu}_{SI}) = 1/\sqrt{1 - FMI}$. Figure 2 of Lewis et al. (2014) plots the standard error ratios against the rates of missingness for a collection of race estimates from the multiply imputed NAMCS data. Their discussion notes no clear trend in the plot, and attributes that to the variability of design effects across the different estimates. Our Figure 1 plots the ratios as a function of missingness rates across different design effects ($deff_{com}$) based on Approximation (4), but with a different plotting symbol for each designeffect (symbols A through F correspond to design effects 1, 2, 5, 10, 20, and 40,

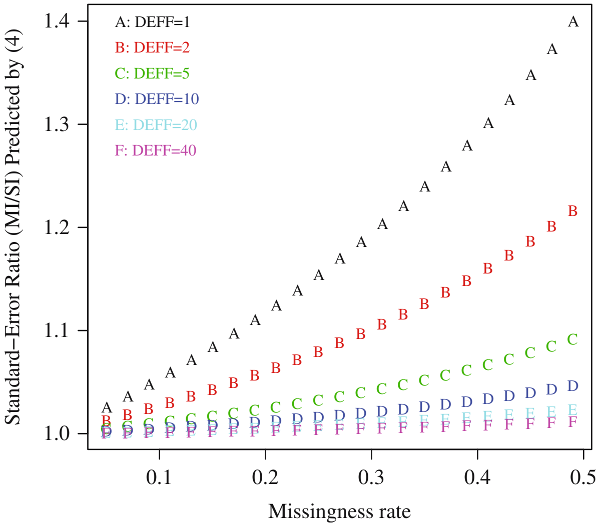


Fig. 1. The Relationship between the Missingness Rate and Standard-Error Ratio (between MI and SI) across Different Completed-Data Design Effects. The symbols A through F correspond to design effects 1, 2, 5, 10, 20, 40, respectively

respectively). For a fixed design effect, the ratio increases as the rate of missingness increases. However, when the design effect is large, the rate of increase of ratios diminishes.

In addition, Figure 3 of Lewis et al. (2014) plots the ratios against the corresponding design effects of the same collection of estimates, clearly showing an inverse relationship between the ratios and design effects. Correspondingly, our Figure 2 plots the ratio as a function of design effect across various missingness rates based on Approximation (4) (symbols A through D correspond to 5%, 10%, 20%, 30% nonresponse rates, respectively). The pattern shown from the actual estimates (Figure 3 of Lewis et al. 2014) is well mimicked here in our Figure 2: as the design effect increases, the ratios decrease and approach 1 across different missingness rates.

Approximations (3) and (4) are based on the simple Model (1) under MCAR, and we only consider the effect of intraclass correlation. The design effects from real complex surveys can be affected by other factors such as unequal weighting, stratification, and multistage sample selection. They can also be affected if the missingness mechanism is more complicated than MCAR. To assess how well the simple approximations work, we predict the ratio of standard errors using Approximations (3) and (4) and compare them with the actual estimates from the NAMCS 2008 data.

The Appendix of Lewis et al. (2014) lists the estimated standard error ratios and design effects from the MI analysis, as well as the nonresponse rates for a wide variety of race estimates. We plug the design effects and nonresponse rates into Approximation (4) and plot the predictions against the actual ratios in the left panel of Figure 3, which also includes a 45-degree line. If the approximations work well, then we would expect to see points clustered around the 45-degree line. It appears that the prediction is reasonable overall and better with smaller standard-error ratios, which likely correspond to estimates

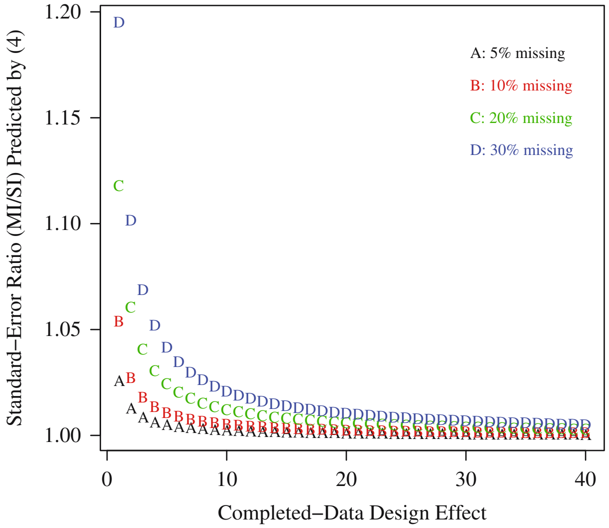


Fig. 2. The Relationship between the Completed-Data Design Effect and Standard-Error Ratio (between MI and SI) across Different Missingness Rates. The symbols A through E correspond to P_{mis} values of 5%, 10%, 20%, and 30%, respectively

with large design effects. On the other hand, Approximation (4) works less well with smaller design effects and tends to underpredict the actual ratios. We surmise that estimates with smaller design effects are likely associated with smaller intracluster correlations, and thus the effects of other factors on the design effect cannot be simply ignored, as they are in the derivation of Approximation (4).

Furthermore, we obtain the design-effect estimates from the observed cases, plug them into Approximation (3), and plot the predicted standard error ratios against the actual ratios in the right panel of Figure 3. As noted at the end of Subsection 3.1, Approximations (3) and (4) can behave differently in more complex situations than assumed in their

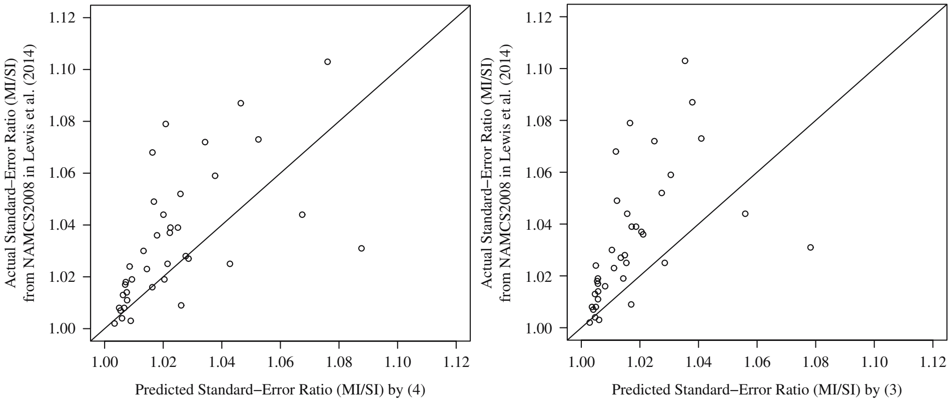


Fig. 3. The Comparison between Actual and Predicted Standard-Error Ratios. Left Panel: by Approximation (4). Right Panel: by Approximation (3)

derivations. Correspondingly, in this example Approximation (3) performs worse than Approximation (4), showing a more severe underprediction for estimates with smaller design effects.

3.3. Practical Implications

From an analyst's/imputer's perspective, Approximations (3) and (4) have simple forms and therefore can be practically useful in exploratory analyses, given the fact that nonresponse rates are readily available and design effects from estimates in complex surveys can be easily estimated from survey statistical packages such as SUDAAN (www.rti.org). For example, before carrying out the combining step in an MI analysis, the analyst might use a singly imputed dataset to obtain point estimates for the estimands of interests and adjust their variance using Approximation (4). Even before conducting MI, an imputer might use Approximation (3) to assess the increase of variance due to MI using design-effect estimates obtained from the observed data. However, we emphasize that the use of these approximations cannot replace principled analyses of missing data (e.g., carefully planned MI and analyses as in Lewis et al. 2014 and other literature).

From a statistical agency's perspective, we recommend releasing FMI estimates for variables with considerable missingness. This is in line with Wagner (2010), which proposed to use FMI as an alternative to the nonresponse rate in data publishing. Despite the common belief that multiply imputed data should be released for public use, we note that releasing only singly imputed data still exists in practice. This might be due in part to limited resources for data production and maintenance, as well as challenges encountered in conveying the concepts of multiple imputation to practical data users (Lewis et al. 2014). Even if multiply imputed data are released, the typical number of data copies (e.g., $D = 5$ or 10) might not be suitable if the FMI is relatively high in certain scenarios (Graham et al. 2007). Therefore, one approach would be to release multiply imputed data with a manageable number of copies to minimize the burden on resources. To compensate for the fact that these numbers might be low in certain cases, the data release could be augmented with the FMI estimates, which are obtained from a much larger D to ensure their accuracy (Harel 2007). The computational burden in obtaining such FMI estimates would be expected to be minimal with current MI software packages. Data users might be able to decide if the number of imputations released are adequate for their analyses of interest given the FMI estimate, for example, by using Rubin (1987, table 4.1) and Graham et al. (2007).

4. Discussion

In this article, we use a one-stage equal clustering sampling design and its model-based characterization to derive the variance components of the MI estimator for the mean estimand. We show that the increase in variance due to MI (or the fraction of missing information) is affected in opposite directions by the frequency of missingness and design effect. Our research is a complement to the empirical investigation in Lewis et al. (2014), one of the first studies identifying such a pattern in practice. Approximations (3) and (4) might be used as simple rules of thumb to gauge the effect of design effects on MI variance estimation.

Approximations (3) and (4) are derived assuming the number of imputations $D \rightarrow \infty$. With a finite D, we conjecture that the main pattern still holds. To see that, note that

$$FMI_D = \frac{r_D + 2/(\nu_D + 3)}{r_D + 1}, \tag{5}$$

where FMI_D defines the fraction of missing information with a finite $D, r_D = (1 + D^{-1})B_D/\bar{W}_D$ and ν_D denotes the degrees of freedom in MI analysis (Rubin 1987; Barnard and Rubin 1999). As $D \rightarrow \infty$, FMI_D approaches FMI which is used in our derivation (Section 3). It could be cumbersome to plug the expressions for variance components (see Appendix) into Equation (5). On the other hand, we can gauge the impact of the design effect with finite D using a well-established large-sample result (Rubin 1987, 114): $V(\bar{Q}_D) = (1 + FMI/D)V(\bar{Q}_\infty)$, which states that the efficiency of the finite-D repeated-imputation estimator relative to the fully efficient infinite-D repeated-imputation estimator is $(1 + FMI/D)^{-1/2}$ in units of standard errors. In our scenario, we show that $FMI \rightarrow 0$ as the design effect increases. This implies that $V(\bar{Q}_D) \approx (1 + 0/D)V(\bar{Q}_\infty) = V(\bar{Q}_\infty)$ accordingly. Therefore the behavior of $V(\bar{Q}_D)$ is expected to be similar to that $V(\bar{Q}_\infty)$ with an increasing design effect.

In addition, one of the key assumptions behind the MI combining rules is that the variance of the within-imputation variance estimator is (asymptotically) much less than the between-imputation variance (Rubin 1987, 89, eq. 3.3.3). That is, $Var(\hat{W}^{(d)}) \ll E(B)$, where $\hat{W}^{(d)}$ is computed from the dth completed dataset. Note that in the scenario considered in this article, as the design effect increases, $FMI \rightarrow 0$, implying that $E(B) \rightarrow 0$. Therefore we believe that using a singly imputed dataset can reliably estimate the within-imputation variance with moderate or large sample size. Obviously using $W = \frac{\sum_{d=1}^{D} \hat{W}^{(d)}}{D}$ (the average from multiply imputed datasets) would produce a more precise estimate for the within-imputation variance. However, this improvement might be minimal compared to the magnitude of the between-imputation variance. More importantly, the main need of multiply imputed data is to reliably estimate the between-imputation variance.

There are several limitations to the current research. First, the derivation assumes MCAR, which can be unrealistic. As a follow-up study to Lewis et al. (2014), the focus of this article is to elucidate the effect of clustering alone on MI variance estimation. More generally, this work can be treated as an extension of Rubin and Schenker (1986), which also focused on MCAR, to clustered data. Assuming a more plausible MAR mechanism implies accounting for the effect of predictive covariates. It is usually believed that FMI would be reduced (i.e., be less than P_{mis}) if the imputation model contained predictive covariates. However, in our limited experience, an explicit formula/relationship has not been proposed and is presumably more complicated. We are currently working on this problem.

Secondly, we conduct the derivations under a rather simplified design (model). The original NAMCS sample design involves features such as stratification and multistage sampling, leading to variable analysis weights which can also affect the design effects (Valliant et al. 2013, sec. 14.4.1 and references therein). In future research, we will study the effect of the design effect on MI estimator with unequal weighting schemes and other factors involved in complex surveys. For example, we might consider a population model

(Valliant et al. 2013, 364)

$$y_{hi} \sim N\left(\mu_h, \sigma_h^2\right), \tag{6}$$

$$P(h=1) = P_1, \ldots, P(h=H) = P_H, \sum_{h=1}^{H} P_h = 1, \tag{7}$$

where h indicates the hth stratum (or poststratum), $i = 1, \ldots, n_h$ indicates the sample selected from that stratum, and P_h indicates the population fraction of the hth stratum. Under such a model, the population mean is $u = \sum_{h=1}^{H} P_h \mu_h$. Unequal weighting occurs when the P_hs are not all equal. We also aim to further extend our work to a more general scenario including both unequal weighting and clustering, understanding how they jointly affect the multiple-imputation variance estimation.

The current research only focuses on the population mean estimand, yet many other estimands such as regression coefficients (controlling for some covariates) are also of major interest in MI analyses. Design effects for regression coefficients have recently been studied (Lohr 2014), and thus it is of interest to include regression analyses in future studies. Furthermore, we will consider extensions to noncontinuous variables, noting that in NAMCS 2008 race is a categorical variable.

Although MI was originally proposed to handle survey nonresponse problems and has been readily applied to a wide variety of data types, systematic methods studies are lacking for understanding its behavior when applied to data with complex survey designs. Together with Lewis et al. (2014), this study can be viewed as a building block for research in this important area. In addition, further studies involving real data, such as that discussed in Lewis et al. (2014), will be invaluable for suggesting theoretical research as well as calibrating it to the real world.

Appendix

MI when τ^2 and σ^2 are known

We first consider the MI scheme which assumes that τ^2 and σ^2 are known. Rewrite Model (1) as

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ \mu_i &\overset{i.i.d.}{\sim} N(\mu, \tau^2) \\ \epsilon_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2). \end{aligned} \tag{1}$$

Also let the cluster-level mean (from observed data) $\bar{y}_{i\cdot,obs} = \frac{\sum_{j=1}^{r} y_{ij}}{r}$; then $Var(\bar{y}_{i\cdot,obs}) = \tau^2 + \frac{\sigma^2}{r}$.

For simplicity, we impose a diffuse prior for μ, i.e., $p(\mu) \propto 1$. The Bayesian imputation scheme consists of

Step 1: Drawing $\mu_i^* \sim p(\mu_i | y_{obs}, \tau^2, \sigma^2)$
Step 2: Drawing $\epsilon_{ij}^* \sim N(0, \sigma^2)$
Step 3: Imputing $y_{ij}^* = \mu_i^* + \epsilon_{ij}^*$ for $i = 1, \ldots, m$ and $j = r+1, \ldots, n$.

First we establish the distributions of the multiply imputed data. After some algebra, μ_i^* in Step 1 can be expressed as $\mu_i^* = \frac{\tau^2 \bar{y}_{i\cdot,obs} + \frac{\sigma^2}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^2 + \frac{\sigma^2}{r}} + a^* + b_i^*$, where $a^* \sim N\left(0, \frac{1}{m}\frac{(\sigma^2/r)^2}{\tau^2+\sigma^2/r}\right)$ and $b_i^* \sim N\left(0, \frac{\tau^2\sigma^2/r}{\tau^2+\sigma^2/r}\right)$. Thus, the imputed value in Step 3 can be expressed as

$$y_{ij}^* = \frac{\tau^2 \bar{y}_{i\cdot,obs} + \dfrac{\sigma^2}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^2 + \dfrac{\sigma^2}{r}} + a^* + b_i^* + \epsilon_{ij}^*, \tag{2}$$

for $i = 1, \ldots, m$ and $j = r + 1, \ldots, n$, where $a^* \perp \{b_i^*\} \perp \{\epsilon_{ij}^*\}$, and $\perp$ indicates independence. Note that a^* is identical across all is whereas the b_i^*s are different across is.

Note that when $\tau^2 \gg \frac{\sigma^2}{r}$, $\frac{\tau^2 \bar{y}_{i\cdot,obs} + \frac{\sigma^2}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^2 + \frac{\sigma^2}{r}} \approx \bar{y}_{i\cdot,obs}$, $a^* \approx 0$, $b_i^* \dot{\sim} N\left(0, \frac{\sigma^2}{r}\right)$, and the distribution of the imputed value can be approximated as $y_{ij}^* \dot{\sim} N\left[\bar{y}_{i\cdot,obs}, \left(1 + \frac{1}{r}\right)\sigma^2\right]$.

By repeating the above process independently D times, we create D completed datasets in which the missing data are imputed as

$$y_{ij}^{*(d)} = \frac{\tau^2 \bar{y}_{i\cdot,obs} + \dfrac{\sigma^2}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^2 + \dfrac{\sigma^2}{r}} + a^{*(d)} + b_i^{*(d)} + \epsilon_{ij}^{*(d)}, \tag{3}$$

for $d = 1, \ldots, D$. Note that $a^{*(d_1)} \perp a^{*(d_2)}$, $\{b_i\}^{*(d_1)} \perp \{b_i\}^{*(d_2)}$, and $\left\{\epsilon_{ij}^{*(d_1)}\right\} \perp \left\{\epsilon_{ij}^{*(d_2)}\right\}$ for any $d_1, d_2 \in (1, \ldots, D)$ and $d_1 \neq d_2$.

Secondly, we derive the forms of mean and variance estimates using Equations (1) and (3). For the dth completed dataset $\left\{\{y_{ij}\}, \left\{y_{ij}^{*(d)}\right\}\right\}$, its mean is

$$\hat{\mu}_{com}^{(d)} = \frac{\sum_{i=1}^m \sum_{j=1}^r y_{ij} + \sum_{i=1}^m \sum_{j=r+1}^n y_{ij}^{*(d)}}{mn} \tag{4}$$

$$= \bar{y}_{\cdot\cdot,obs} + \frac{n-r}{n}\left(a^{*(d)} + \bar{b}^{*(d)} + \bar{\epsilon}_{\cdot\cdot}^{*(d)}\right) \tag{5}$$

where $\bar{b}^{*(d)} = \frac{\sum_{i=1}^m b_i^{*(d)}}{m}$, and $\bar{\epsilon}_{\cdot\cdot}^{*(d)} = \frac{\sum_{i=1}^m \sum_{j=r+1}^n \bar{\epsilon}_{ij}^{*(d)}}{m(n-r)}$.

The MI estimator is $\hat{\mu}_{MI} = \frac{\sum_{d=1}^D \hat{\mu}_{com}^{(d)}}{D}$. Because the $\hat{\mu}_{com}^{(d)}$s are identically distributed (yet correlated with each other), we have

$$E(\hat{\mu}_{MI}) = E\left(\hat{\mu}_{com}^{(d)}\right) = E(\bar{y}_{\cdot\cdot,obs}) = \mu; \tag{6}$$

thus μ_{MI} is unbiased.

Its variance is $Var(\hat{\mu}_{MI}) = Var\left(\frac{\sum_{d=1}^D \hat{\mu}_{com}^{(d)}}{D}\right) = \frac{D-1}{D} cov\left(\hat{\mu}_{com}^{(1)}, \hat{\mu}_{com}^{(2)}\right) + \frac{1}{D} Var\left(\hat{\mu}_{com}^{(1)}\right)$. Let $D \rightarrow \infty$; then $Var(\hat{\mu}_{MI}) \rightarrow cov\left(\hat{\mu}_{com}^{(1)}, \hat{\mu}_{com}^{(2)}\right) = cov\left[\bar{y}_{\cdot\cdot,obs} + \frac{n-r}{n}\left(a^{*(1)} + \bar{b}^{*(1)} + \bar{\epsilon}_{\cdot\cdot}^{*(1)}\right),\right.$ $\left.\bar{y}_{\cdot\cdot,obs} + \frac{n-r}{n}\left(a^{*(2)} + \bar{b}^{*(2)} + \bar{\epsilon}_{\cdot\cdot}^{*(2)}\right)\right]$. After some algebra, we have as $D \rightarrow \infty$

$$Var(\hat{\mu}_{MI}) \rightarrow \tau^2/m + \frac{\sigma^2}{mr} = Var(\hat{\mu}_{obs}). \tag{7}$$

We now study the between-imputation variance $B = \frac{\sum_{d=1}^D \left(\hat{\mu}_{com}^{(d)} - \hat{\mu}_{MI}\right)^2}{D-1}$. Again, because the $\hat{\mu}_{com}^{(d)}$s are identically distributed, $E(B) = \frac{D}{D-1}\left[Var\left(\hat{\mu}_{com}^{(1)}\right) - Var(\hat{\mu}_{MI})\right]$. As $D \rightarrow \infty$, $E(B) \rightarrow Var\left(\hat{\mu}_{com}^{(1)}\right) - Var(\hat{\mu}_{MI})$. In addition, $Var\left(\hat{\mu}_{com}^{(1)}\right) = Var(\bar{y}_{\cdot\cdot,obs}) + \left(\frac{n-r}{n}\right)^2$

$\left[Var(a^{*(1)}) + Var(\bar{b}^{*(1)}) + Var\left(\bar{\epsilon}^{*(1)}_{\cdot\cdot}\right)\right]$. Plugging in $Var(a^{*(1)}) = \frac{1}{m}\frac{(\sigma^2/r)^2}{\tau^2+\sigma^2/r}$, $Var(\bar{b}^{*(1)}) = \frac{1}{m}\frac{\tau^2\sigma^2/r}{\tau^2+\sigma^2/r}$, and $Var\left(\bar{\epsilon}^{*(1)}_{\cdot\cdot}\right) = \frac{\sigma^2}{m(n-r)}$, we obtain $Var\left(\hat{\mu}^{(1)}_{com}\right) = \frac{\tau^2}{m} + 2\frac{\sigma^2}{mr} - \frac{\sigma^2}{mn}$. Thus as $D \longrightarrow \infty$,

$$E(B) \longrightarrow \frac{\sigma^2}{mr} - \frac{\sigma^2}{mn}. \tag{8}$$

For the dth completed dataset, the within-imputation variance $W^{(d)}$ is calculated as $\frac{\sum_{i=1}^{m}\left(\bar{y}^{(d)}_{i\cdot,com} - \bar{y}^{(d)}_{\cdot\cdot,com}\right)^2}{m(m-1)}$. The average of the within-imputation variance is $W = \frac{\sum_{(d)=1}^{D} W^{(d)}}{D}$. Because the $W^{(d)}$s are identically distributed across ds, $E(W) = E(W^{(d)})$. For simplicity, we ignore the notational index d in the following derivations. Note that $E(W) = E\left(\frac{\sum_{i=1}^{m}(\bar{y}_{i\cdot,com} - \bar{y}_{\cdot\cdot,com})^2}{m(m-1)}\right) = E\left(\frac{\bar{y}^2_{1\cdot,com} - \bar{y}^2_{\cdot\cdot,com}}{m-1}\right)$ given the identical distributions of $\bar{y}_{i\cdot,com}$ across is. In addition, $E(\bar{y}_{i\cdot,com}) = E(\bar{y}_{\cdot\cdot,com}) = \mu$, and $Var(\bar{y}_{\cdot\cdot,com}) = \frac{1}{m}Var(\bar{y}_{1\cdot,com}) + \frac{m-1}{m}cov(\bar{y}_{1\cdot,com}, \bar{y}_{2\cdot,com})$. Therefore $E(W) = \frac{Var(\bar{y}_{1\cdot,com}) - Var(\bar{y}_{\cdot\cdot,com})}{m-1} = \frac{Var(\bar{y}_{1\cdot,com}) - cov(\bar{y}_{1\cdot,com}, \bar{y}_{2\cdot,com})}{m}$. After some algebra, we obtain that $Var(\bar{y}_{1\cdot,com}) - cov(\bar{y}_{1\cdot,com}, \bar{y}_{2\cdot,com}) = \left(\frac{r}{n} + \frac{n-r}{n}\frac{\tau^2}{\tau^2+\sigma^2/r}\right)^2 Var(\bar{y}_{1\cdot,obs}) + \left(\frac{n-r}{n}\right)^2\left[Var\left(b^*_1\right) + Var\left(\bar{\epsilon}^*_{1\cdot}\right)\right]$. Plugging in $Var(\bar{y}_{1\cdot,obs}) = \tau^2 + \sigma^2/r$, $Var(\bar{y}_{\cdot\cdot,obs}) = \frac{1}{m}(\tau^2 + \sigma^2/r)$, $Var\left(b^*_1\right) = \frac{\tau^2\sigma^2/r}{\tau^2+\sigma^2/r}$, and $Var\left(\bar{\epsilon}^*_{1\cdot}\right) = \sigma^2/(n-r)$, we obtain

$$E(W) = \frac{\tau^2}{m} + \frac{\sigma^2}{mn}. \tag{9}$$

Based on (7), (8), and (9), we have $E(W) + E(B) \longrightarrow \tau^2/m + \sigma^2/mr = Var(\hat{\mu}_{MI})$ as $D \longrightarrow \infty$, consistent with Rubin's variance combination formulae.

MI when τ^2 and σ^2 are unknown

More realistically, suppose MI is conducted without knowing τ^2 and σ^2. We impose proper prior distributions for these parameters: $p(\mu) \sim N\left(0, \sigma^2_\mu\right)$, $p(\tau^2) \sim IG(A_{\tau^2}, B_{\tau^2})$, and $p(\sigma^2) \sim IG(A_{\sigma^2}, B_{\sigma^2})$, where IG denotes the inverse-gamma distribution. These priors are often employed in hierarchical Bayesian models (Gelman et al. 2004).

The variance components and imputations are drawn from an integrated Gibbs sampling algorithm sketched as follows:

Step 1: Draw τ^{*2} from $p(\tau^2|y_{obs}, \mu_i, \sigma^2)$;
Step 2: Draw σ^{*2} from $p(\sigma^2|y_{obs}, \mu_i, \tau^2)$;
Step 3: Draw μ^*_i from $p(\mu^*_i|y_{obs}, \tau^2, \sigma^2)$.

For a single imputation, we repeat Steps 1–3 until the Gibbs chain converges. We then draw $\epsilon^*_{ij} \sim N(0, \sigma^{*2})$, and impute $y^*_{ij} = \mu^*_i + \epsilon^*_{ij}$ for $i = 1, \ldots, m$ and $j = r+1, \ldots, n$, where μ^*_i and σ^{*2} are the draws from the last iteration of the chain. We repeat this procedure independently D times to construct D completed datasets.

The posterior distributions of τ^2 and σ^2 under a common class of priors (including ours here) are very complicated (Box and Tiao 1973, chap. 6), and therefore it is difficult to obtain their moments using explicit formulaes. Nevertheless, we can assess the MI variance estimators asymptotically. In a general scenario, as stated in (Gelman et al. 2004, 587), the posterior distribution of a parameter θ approaches normality with mean θ_0 and variance $[nJ(\theta_0)]^{-1}$ as the sample size $n \longrightarrow \infty$ and subject to some regularity conditions, where θ_0 is the value that minimizes the Kullback-Leibler information and J is the Fisher information. Therefore $E(\theta^*) \longrightarrow \theta_0$ and $E(\theta^{*2}) \longrightarrow \theta_0^2$ as $n \longrightarrow \infty$, where θ^* is a draw from the posterior distribution of θ. In our context, θ can be viewed as a smooth function, say f, of τ^2 and σ^2, θ_0 as the same function evaluated at the true τ^2 and σ^2 (from the frequentist's perspective), and sample size n can be viewed as the number of clusters m. Correspondingly, we have $E[f(\tau^{*2}, \sigma^{*2})] \longrightarrow f(\tau^2, \sigma^2)$ and $E[f(\tau^{*2}, \sigma^{*2})^2] \longrightarrow f^2(\tau^2, \sigma^2)$ as $m \longrightarrow \infty$, where τ^{*2} and σ^{*2} are draws from the posterior distributions of τ^2 and σ^2.

After the Gibbs sampler converges, imputation y_{ij}^* can be expressed as

$$y_{ij}^* = \frac{\tau^{*2}\bar{y}_{i\cdot,obs} + \dfrac{\sigma^{*2}}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^{*2} + \dfrac{\sigma^{*2}}{r}} + a^* + b_i^* + \epsilon_{ij}^* \tag{10}$$

for $i = 1, \ldots, m$ and $j = r + 1, \ldots, n$, and $a^* \perp \{b_i^*\} \perp \{\epsilon_{ij}^*\}$, where $a^* \sim N\left(0, \frac{1}{m}\frac{(\sigma^{*2}/r)^2}{\tau^{*2}+\sigma^{*2}/r}\right)$, and $b_i^* \overset{i.i.d.}{\sim} N\left(0, \frac{\tau^{*2}\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right)$, and $\epsilon_{ij}^* \overset{i.i.d.}{\sim} N(0, \sigma^{*2})$. Here τ^{*2} and σ^{*2} are draws from their posterior distributions $p(\tau^2, \sigma^2 | y_{obs})$.

Similar to the case in which τ^2 and σ^2 are known, it is not hard to show that

$$E(\hat{\mu}_{MI}) = E(\hat{\mu}_{com}^{(d)}) = \mu. \tag{11}$$

As $D \longrightarrow \infty$,

$$Var(\hat{\mu}_{MI}) \longrightarrow \frac{\tau^2}{m} + \frac{\sigma^2}{mr} = Var(\bar{y}_{\cdot\cdot,obs}). \tag{12}$$

Now, the between-imputation variance is computed as $B = \frac{\sum_{d=1}^{D}\left(\hat{\mu}_{com}^{(d)} - \hat{\mu}_{MI}\right)^2}{D-1}$. Again, because the $\hat{\mu}_{com}^{(d)}$s are identically distributed, $E(B) = \frac{D}{D-1}\left[Var\left(\hat{\mu}_{com}^{(1)}\right) - Var(\hat{\mu}_{MI})\right] \longrightarrow Var\left(\hat{\mu}_{com}^{(1)}\right) - Var(\hat{\mu}_{MI})$ as $D \longrightarrow \infty$. In addition, $Var\left(\hat{\mu}_{com}^{(1)}\right) = Var\left[\bar{y}_{\cdot\cdot,obs} + \frac{n-r}{n}\left(a^{*(1)} + \bar{b}^{*(1)} + \bar{\epsilon}_{\cdot\cdot}^{*(1)}\right)\right] = Var(\bar{y}_{\cdot\cdot,obs}) + \left(\frac{n-r}{n}\right)^2 \left[Var(a^{*(1)}) + Var(\bar{b}^{*(1)}) + Var\left(\bar{\epsilon}_{\cdot\cdot}^{*(1)}\right)\right]$. More specifically, $Var(a^{*(1)}|\tau^2, \sigma^2) = Var[E(a^{*(1)}|\tau^{*2}, \sigma^{*2})] + E[Var(a^{*(1)}|\tau^{*2}, \sigma^{*2})] = 0 + E\left(\frac{1}{m}\ \frac{(\sigma^{*2}/r)^2}{\tau^{*2}+\sigma^{*2}/r}\right) \longrightarrow \frac{1}{m}\ \frac{(\sigma^2/r)^2}{\tau^2+\sigma^2/r}$ as $m \longrightarrow \infty$. The last convergence holds because of the aforementioned Bayesian asymptotic arguments. Similarly, we have $Var(\bar{b}^{*(1)}|\tau^2, \sigma^2) = E\left(\frac{1}{m}\ \frac{\tau^{*2}\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right) \longrightarrow \frac{1}{m}\ \frac{\tau^2\sigma^2/r}{\tau^2+\sigma^2/r}$, and $Var\left(\bar{\epsilon}_{\cdot\cdot}^{*(1)}|\tau^2, \sigma^2\right) = E\left(\frac{\sigma^{*2}}{m(n-r)}\right) \longrightarrow \frac{\sigma^2}{m(n-r)}$. This leads to

$$E(B) \longrightarrow \frac{\sigma^2}{mr} - \frac{\sigma^2}{mn}. \tag{13}$$

For ease of notation, we drop the conditioning on τ^2 and σ^2 in the following expressions in evaluating the within-imputation variance W and assume $m \longrightarrow \infty$. Similar to the case with τ^2 and σ^2 known, we have $E(W) = \frac{Var(\bar{y}_{1\cdot,com}) - Cov(\bar{y}_{1\cdot,com}, \bar{y}_{2\cdot,com})}{m}$. First, $Var(\bar{y}_{1\cdot,com}) = Var\left[\frac{r}{n}\bar{y}_{1\cdot,obs} + \frac{n-r}{n}\ \frac{\tau^2}{\tau^2+\sigma^2/r}\bar{y}_{1\cdot,obs} + \frac{n-r}{n}\frac{\sigma_r^2\bar{y}_{\cdot\cdot,obs}}{\tau^2+\sigma^2/r} + \frac{n-r}{n}\left(a^* + b_1^* + \bar{\epsilon}_1^*\right)\right] = Var\left[\left(\frac{r}{n} + \frac{n-r}{n}\ \frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}\right] + Var\left[\left(\frac{n-r}{n}\ \frac{\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{\cdot\cdot,obs}\right] + 2*$

$$Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs},\left(\frac{n-r}{n}\frac{\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{\cdot\cdot,obs}\right]$$
$$+\left(\frac{n-r}{n}\right)^2\left[Var(a^*)+Var\left(b_1^*\right)+Var\left(\bar{\epsilon}_{1t}^*\right)\right].$$

In addition, $Cov(\bar{y}_{1\cdot,com},\bar{y}_{2\cdot,com})=Cov\Big[\frac{r}{n}\bar{y}_{1\cdot,obs}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\bar{y}_{1\cdot,obs}+\frac{n-r}{n}\frac{\frac{\sigma^{*2}}{r}\bar{y}_{\cdot\cdot,obs}}{\tau^{*2}+a^{*2}/r}+$ $\frac{n-r}{n}\left(a^*+b_1^*+\bar{\epsilon}_{1\cdot}^*\right),\frac{r}{n}\bar{y}_{2\cdot,obs}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\bar{y}_{2\cdot,obs}+\frac{n-r}{n}\frac{\sigma^{*2}/r\bar{y}_{\cdot\cdot,obs}}{\tau^{*2}+\sigma^{*2}/r}+\frac{n-r}{n}\left(a^*+b_2^*+\bar{\epsilon}_{2\cdot}^*\right)\Big]=$ $Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs},\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}\right]+Var\left[\left(\frac{n-r}{n}\frac{\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{\cdot\cdot,obs}\right]$ $+2Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,}\ obs,\left(\frac{n-r}{n}\frac{\sigma^{*2}/r}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{\cdot\cdot,obs}\right]+\left(\frac{n-r}{n}\right)^2Var(a^*)$. Note that unlike the scenario where τ^2 and σ^2 are known, there exists a covariance between $\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}$ and $\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}$ (induced by τ^{*2} and σ^{*2}) in spite of the independence between $\bar{y}_{1\cdot,obs}$ and $\bar{y}_{2\cdot,obs}$.

Therefore $Var(\bar{y}_{1\cdot,com})-Cov(\bar{y}_{1\cdot,com},\bar{y}_{2\cdot,com})=Var\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}\right]-$ $Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs},\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}\right]+\left(\frac{n-r}{n}\right)^2\left[Var\left(b_1^*\right)+Var\left(\bar{\epsilon}_{1\cdot}^*\right)\right]$. More specifically, $Var\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}\right]=Var\left\{E\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,}\right.\right.$ $\left.\left.obs|\tau^{*2},\sigma^{*2}\right]\right\}+E\left\{Var\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}|\tau^{*2},\sigma^{*2}\right]\right\}=\mu^2Var\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)$ $+(\tau^2+\sigma^2/r)E\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)^2$. Note that the second term $(\tau^2+\sigma^2/r)E$ $\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)^2\rightarrow(\tau^2+\sigma^2/r)\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^2}{\tau^2+\sigma^2/r}\right)^2$.

Furthermore,

$$Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs},\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}\right]$$
$$=Cov\left\{E\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs}|\tau^{*2},\sigma^{*2}\right],\right.$$
$$\left.E\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}|\tau^{*2},\sigma^{*2}\right]\right\}$$
$$+E\left\{Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{1\cdot,obs},\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\bar{y}_{2\cdot,obs}\right]\right\}$$
$$=Cov\left[\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\mu,\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right)\mu\right]$$
$$=\mu^2Var\left(\frac{r}{n}+\frac{n-r}{n}\frac{\tau^{*2}}{\tau^{*2}+\sigma^{*2}/r}\right).$$

Finally, $Var\left(b_1^*\right)\rightarrow\frac{\tau^2\sigma^2/r}{\tau^2+\sigma^2/r}$ and $Var\left(\bar{\epsilon}_{1\cdot}^*\right)\rightarrow\frac{\sigma^2}{n-r}$.

Plugging all these expressions into $E(W)$, we obtain

$$E(W) \rightarrow \tau^2/m + \sigma^2/mn \tag{14}$$

Based on (12), (13), and (14), Rubin's variance formulae are valid asymptotically.

5. References

Andridge, R.R. 2011. "Quantifying the Impact of Fixed Effects Modeling of Clusters in Multiple Imputation for Cluster Randomized Trials." *Biometrical Journal* 53: 57–74. Doi: http://dx.doi.org/10.1002/bimj.201000140.

Barnard, J. and D.B. Rubin. 1999. "Small-Sample Degrees of Freedom With Multiple Imputation." *Biometrika* 86: 949–955.

Box, G.E.P. and G.C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. NewYork: Wiley.

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

Graham, J., A. Olchowsi, and T. 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.

Harel, O. 2007. "Inferences on Missing Information Under Multiple Imputation and Two-Stage Multiple Imputation." *Statistical Methodology* 4: 75–89. Doi: http://dx.doi.org/10.1016/j.stamet.2006.03.002.

Harel, O. and X.H. Zhou. 2007. "Multiple Imputation: Review of Theory, Implementation, and Software." *Statistics in Medicine* 26: 3057–3077. Doi: http://dx.doi.org/10.1002/sim.2787.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. New York: CRC Press.

Kim, J.K. and J. Shao. 2014. *Statistical Methods for Handling Incomplete Data*. Boca Raton: CRC Press.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Lewis, T., E. Goldberg, N. Schenker, V. Beresovsky, S. Schappert, S. Decker, N. Sonnenfeld, and I. Shimizu. 2014. "The Relative Impacts of Design Effects and Multiple 21 Imputation on Variance Estimates: A Case Study With the 2008 National Ambulatory Medical Care Survey." *Journal of Official Statistics* 30: 147–161. Doi: http://dx.doi.org/10.2478/jos-2014-0008.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis of Missing Data*. New York: Wiley.

Lohr, S.L. 2014. "Design Effects for a Regression Slope in a Cluster Sample." *Journal of Survey Statistics and Methodology* 2: 97–125. Doi: http://dx.doi.org/10.1093/jssam/smu003.

Meng, X.L. 1995. "Multiple Imputation With Uncongenial Sources of Input (with discussion)." *Statistical Science* 10: 538–573.

Reiter, J.P., T.E. Raghunathan, and S.K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32: 143–149.

Rubin, D.B. 1978. “Multiple Imputations in Sample Surveys – a Phenomenological Bayesian Approach to Nonresponse.” In Proceedings of the Survey Research Methods Section of the American Statistical Association, date and place, 20–34.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. and N. Schenker. 1986. “Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse.” *Journal of the American Statistical Association* 81: 366–374. Doi: http://dx.doi.org/10.1080/01621459.1986.10478280.

Schafer, J.L. and R.M. Yucel. 2002. “Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values.” *Journal of Computational and Graphical Statistics* 11: 421–442. Doi: http://dx.doi.org/10.1198/106186002760180608.

Schenker, N., T.E. Raghunathan, P.L. Chiu, D.M. Makuc, G. Zhang, and A.J. Cohen. 2006. “Multiple Imputation of Missing Income Data in the National Health Interview Survey.” *Journal of the American Statistical Association* 101: 924–933. Doi: http://dx.doi.org/10.1198/016214505000001375.

Skinner, C.J., D. Hold, and T.F.M. Smith. 1989. *Analysis of Complex Surveys*. West Sussex: Wiley.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

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

Van Buuren, S. 2012. *Flexible Imputation for Missing Data*. Boca Raton: CRC Press.

Wagner, J. 2010. “The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data”, *Public Opinion Quarterly* 74, 233–243. DOI: http://www.dx.doi.org/10.1093/poq/nfq007.

Received October 2014
Revised June 2015
Accepted June 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 165–186, http://dx.doi.org/10.1515/JOS-2016-0008

Sample Representation and Substantive Outcomes Using Web With and Without Incentives Compared to Telephone in an Election Survey

Oliver Lipps[1] *and Nicolas Pekari*[2]

The objective of this article is to understand how the change of mode from telephone to web affects data quality in terms of sample representation and substantive variable bias. To this end, an experiment, consisting of a web survey with and without a prepaid incentive, was conducted alongside the telephone Swiss election survey. All three designs used identical questionnaires and probability samples drawn from a national register of individuals.

First, our findings show that differences in completion rates mostly reflect different levels of coverage in the two modes. Second, incentives in the web survey strongly increase completion rates of all person groups, with the exception of people without Internet access or limited computer literacy. Third, we find voting behavior to be much closer to official figures in the web with the incentive version compared to the two other designs. However, this is partly due to the different sociodemographic compositions of the samples. Other substantive results suggest that the incentive version includes harder-to-reach respondents. Unit costs are much lower in the two web designs compared to the telephone, including when a relatively high incentive is used. We conclude that in countries with high Internet penetration rates such as Switzerland, web surveys are already likely to be highly competitive.

Key words: Web surveys; mode experiment; incentive effects; individual register frame; national election survey.

1. Nonobservation and Measurement Issues in Web and Telephone Surveys

1.1. Nonobservation and Sample Composition Differences

In the early 2000s, survey methodologists began studying web surveys as an alternative to traditional survey modes. From the outset, the major concern was the limited Internet access of some groups regarding various demographic variables such as sex and age (Brandtzæg et al. 2011; Zickhur and Smith 2012), education (Struminskaya et al. 2014; Mohorko et al. 2013a), marital status (Struminskaya et al. 2014), and urbanicity (Mohorko et al. 2013a). While in these studies those with Internet at home are more often male, young, highly educated, unmarried, and live in more urbanized areas, Revilla and Saris (2013) for instance find very little difference between the face-to-face European Social Survey (ESS) and the Longitudinal Internet Studies for the Social Sciences (LISS) online

[1] Corresponding author: FORS – Swiss Centre of Expertise in the Social Sciences, c/o University of Lausanne Quartier Mouline Lausanne 1015, Lausanne, Switzerland. Email: oliver.lipps@fors.unil.ch
[2] FORS – Swiss Centre of Expertise in the Social Sciences, c/o University of Lausanne Quartier Mouline Lausanne 1015, Lausanne, Switzerland. Email: nicolas.pekari@fors.unil.ch

panel in the Netherlands. The latter does make a strong effort to improve representativity, including providing a computer with an Internet connection to those who lack one, but it shows the potential for high-quality surveys using the web. The trend of increasing Internet availability, with a penetration rate as high as 86.6% in Switzerland in 2014 and 82.1% in Europe 2015 (International Telecommunication Union, http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx), will also continue to work in favor of web surveys in terms of an increasingly smaller substantive variable bias, with a smaller bias in more economically developed countries (Mohorko et al. 2013a).

Conversely, in telephone surveys, the increasing landline telephone undercoverage in Europe is causing an increasing bias in both demographic and substantive variables (Mohorko et al. 2013b). For most Western European countries, the number of households without a telephone plays a minor role (Busse and Fuchs 2012). This is due to an increasing proportion of households who have substituted a mobile phone for their landline (Ernst Stähli 2012; Joye et al. 2012; Blumberg and Luke 2013; Link and Fahimi 2008), and those with an unlisted landline number (Von der Lippe et al. 2011). In Switzerland, the Swiss Federal Statistical Office (SFSO) recorded an average matching rate of individuals sampled at random from their individual population register and matched against their own register of landline numbers of around 76% (Joye 2012). The SFSO does not provide unlisted landline numbers to other organizations.

Those without a landline differ in particular from those with a listed number. People without a landline are younger (Joye et al. 2012; Blumberg and Luke 2013), live in larger municipalities, and are more often unmarried (Von der Lippe et al. 2011; Lipps and Kissau 2012) or of foreign nationality (Cobben and Bethlehem 2005; Lipps and Kissau 2012). In addition, the increasing volume of marketing calls is causing certain groups of potential telephone-survey respondents, especially the more affluent, to employ gatekeeper technologies to screen calls. These issues render the telephone increasingly problematic as a survey tool in terms of achieving a representative sample of a population.

A number of studies have compared web with telephone surveys (Fricker et al. 2005; Braunsberger et al. 2007; Chang and Krosnick 2009; Dillman et al. 2009; Nagelhout et al. 2010; Yeager et al. 2011). However, most of the web surveys considered do not use a probability sample. This is unfortunate as, according to the current state of research, nonprobability web surveys might not be a viable alternative in terms of their representativity (Chang and Krosnick 2009; Baker et al. 2010; Yeager et al. 2011). Because representative lists of e-mail addresses of the general population are not available, most web surveys use a sample drawn from an Random Digital Dialing (RDD)-screened sample (Chang and Krosnick 2009; Yeager et al. 2011) or from online panels (Braunsberger et al. 2007; Baker et al. 2010; Nagelhout et al. 2010). Undercoverage, selection effects and panel conditioning might bias inferences drawn from such samples (Warren and Halpern-Manners 2012), including estimates of population frequencies and percentages, and often remain a neglected problem (Schonlau et al. 2009). Other studies use subsamples with Internet access drawn from an RDD sample, which are then randomly assigned to telephone or web (Fricker et al. 2005). One exception is Atkeson et al. (2014), who compare a probability-based telephone poll and a web poll and find only a few differences between survey modes regarding demographic characteristics. One data source that allows for a comparison of web and telephone respondents, under the premise that face-to-face respondents are representative of the population, is the ESS 2010. Among respondents of the ESS 2010, 24% of Swiss adults never use the Internet and ten percent do not own a landline telephone. Our calculations reveal that older, female, married people, those living in larger households, the Swiss-German-speaking part of Switzerland, or in smaller municipalities are more likely to possess a landline. Internet use is more common among younger individuals, men, or those living in larger households. With respect to substantive variables, controlled for sociodemography, we find a positive correlation between Internet use and both political interest and voting behavior, but no correlation of the latter variables with landline possession.

Unlike undercoverage, differences in nonresponse between the two modes are not directly comparable because this depends strongly both on sample matching effort in the case of the telephone (if the sample was drawn from a source that does not include telephone numbers; for the survey used in this article, see Lipps et al. 2015) and on fieldwork effort. In addition, by optimizing the design for each mode, many factors vary other than the mode of interview (Dillman 2000; 2011). Examples include sample members contacted through different methods, or a difference in the number of calls or the number of reminders, as well as the use of incentives. Despite the difficulties of comparing response rates and respondent compositions across different modes, some authors have attempted this, and there seems to be some agreement in that response rates to probability-based web surveys are lower than those for comparable interviewer-based surveys. For example, one meta-analysis finds that web surveys on average yield an 11% lower response rate compared to other modes (Lozar Manfreda et al. 2008). A more recent example is that of Sinclair et al. (2012), who compare web and telephone in a community-based epidemiological survey in Melbourne, Australia, and find an Internet response rate of 4.7% and a telephone response rate of 30.2%. When interpreting these figures, we must bear in mind that response rates are very much dependent on contexts such as the topic, the country, or fieldwork efforts (see e.g., Groves et al. 2004a,b).

Given the issues related to comparing specific effects, we choose to measure a "mode system effect" (Biemer and Lyberg 2003, also Biemer 1988) in order to compare data quality. This means that rather than trying to isolate the effects of mode with respect to the quality indicators we use, we measure effects of the entire data-collection process designed around a specific mode and thus compare whole *systems* of data collection. We measure the difference between the two systems regarding completion rates, sociodemographic composition, distributions of substantive variables, and the cost of data collection. Note that we abstain from using the term "response rate", because this does not include errors arising from noncoverage. We instead talk of nonobservation or completion rate (Callegaro and DiSogra 2008). In addition, we are able to measure bias regarding sociodemographic and political behavior variables by comparing survey outcomes with official figures.

1.2. Measurement Effects

In addition to differences in coverage and nonresponse between phone and web modes, differences also arise because of the mode itself. In particular, interviewer-administered surveys are known to produce more socially desirable answers (Krosnick 1991), whether on the telephone (Kreuter et al. 2008; Chang and Krosnick 2009; Atkeson et al. 2014) or face-to-face (De Leeuw 2005). However, these differences can also be attributed in part to selection effects (Chang and Krosnick 2009).

In addition, in visual modes, such as web-administered questionnaires, respondents tend to think in the order in which the response categories are presented and are more likely to choose those presented at the beginning of a list of response alternatives than those at the end (primacy effect). On the contrary, in an aural format, such as telephone, respondents are expected to wait until the interviewer has read all the response categories and are thus more likely to start thinking about the last alternatives read to them (recency effect; e.g., Christian et al. 2008). De Leeuw (2005), Dillman and Christian (2005) and Dillman et al. (2009) note, however, that the evidence found for this so far is not completely consistent.

Regarding the accuracy of reported political behavior and preferences, Stephenson and Crête (2011) find the number of differences in point estimates to be relatively high. Malhotra and Krosnick (2007), comparing web data with the 2000 and the 2004 American National Election Study (ANES) face-to-face surveys, find accuracy to be higher for the face-to-face probability-sample data than for the online-panel sample data in the majority of the comparisons, in particular for voting turnout and party choice. However, more recent research tends to show that high-quality web surveys are comparable to traditional modes. For instance, Ansolabehere and Schaffner (2014) state that with the correct sampling methods and weighting, web surveys can produce point estimates comparable to a telephone survey.

2. Incentive Effects in Web Surveys

2.1. *Effects on Response Rates*

In addition to high nonresponse rates, partial completes (Peytchev 2009) may be a concern in web surveys. Different types of incentives are thus often used to mitigate these problems and a number of experiments have been designed to analyze the effects of different types of incentives both for surveys that recruit online and offline. Regarding online recruitment, one meta-analysis of 32 surveys showed that incentives significantly increased the motivation to start a survey, and another meta-analysis of 26 studies found that incentives were effective in motivating participants to finish a web survey (Göritz 2006). In addition, studies using online access panels tend to support the reciprocity norm (Groves et al. 1992), in that prepaid incentives are more effective than conditional incentives in web surveys (Bosnjak 2005; Su et al. 2008). In addition, the economic exchange theory (Ryu et al. 2006) is supported because cash (Birnholtz et al. 2004; van Veen et al. 2011) and larger incentives have been found to be more effective (Schaurer et al. 2012). Both effects can also be found in telephone surveys (Singer et al. 2000; Curtin et al. 2007; Sánchez-Fernández et al. 2010). However, in their study Bosnjak and Tuten (2003) find that prepaid incentives do not increase completion rates.

Only a few studies test incentive effects for samples recruited offline. In the study by Alexander et al. (2008), which recruited sample members by mail, the prepaid incentive had better overall enrollment rates. The rates also slightly increased with greater incentive value although men responded more to the $2 bill than either the $1 or the $5 bill. In an address-based experiment, Messer and Dillman (2011) found that a $5 prepaid cash incentive was effective at improving both web and mail response rates. Providing an incentive significantly increased web response rates by 17.9 percentage points. Scherpenzeel and Toepoel (2012) also report positive prepaid incentive effects from an offline (via telephone or face-to-face) recruitment for the Dutch LISS panel survey. However, they do not find increased response rates in the 20 Euro and 50 Euro designs compared to those seen at the 10 Euro level. Finally, Parsons and Manierre (2014) report that the use of unconditional incentives improved response rates in a web survey based on a random sample of college students.

In addition to incentives, Messer and Dillman (2011) tested the use of priority mail, which did not improve response rates. However, the inclusion of an additional $5 incentive in combination with priority mail produced slightly higher response rates, although the differences were not statistically significant.

2.2. *Effects on Sample Composition and Response Quality*

Research on the effects of incentives on sample composition and response quality in web surveys is scarce. Two possible hypotheses are generally advanced: according to the first, the incentive draws respondents who would not have responded otherwise and whose response quality is possibly poorer, and, according to the second, the reward from the incentive leads to an improvement in the quality of answers (Singer and Ye 2013). The results regarding these hypotheses are mixed, and most research has found no significant difference in response quality. Göritz (2004), for instance, reports only small effects. In a meta-analysis, Singer and Bossarte (2006) find that incentives may raise response rates without decreasing the nonresponse bias because they motivate individuals who were already more predisposed to respond. They conclude that "more research is needed on how monetary incentives can reduce nonresponse bias rather than merely raising the rate of response" (413). Parsons and Manierre (2014) in turn find that prepaid cash incentives may actually produce results that are less representative of the target population. In a metastudy, Singer and Ye (2013) report that most studies find no or only small effects on sample compositions and call for additional research on the subject. Regarding the effect of incentives on substantive results, Teisl et al. (2006) find that different incentive conditions yield different responses, even when response rates and demographic compositions are the same, concluding that incentives do draw different kinds of respondents. However, direct effects of incentives on response distributions have not been found (Singer and Ye 2013).

Although this literature review reports some inconsistent findings, we are able to draw several preliminary conclusions:

- With respect to sample composition, respondents to web surveys tend to be younger, male, better educated, unmarried, and live in more urbanized areas. Conversely, respondents to landline surveys tend to be older, less educated, natives, and living in less urban areas.
- Studies that compare differences in substantive variables in interviewer-based and web mode are scarce. Most studies use the face-to-face mode as the interviewer-based survey mode. There is a tendency for face-to-face surveys to achieve more accurate results.

- While unconditional cash incentives are able to increase response rates in web surveys, the effect on the sample composition and on substantive answers is unclear. However, the majority of studies uses samples of special population groups (e.g., students), RDD-screened samples, or online panels.

In our view, the biggest issue with existing studies is that most mode comparisons, including web surveys, are based on nonprobability samples or reference surveys (usually face-to-face) that are biased by nonresponse. In addition, RDD-screened samples suffer from undercoverage problems. As for substantive variables, comparisons using interviewer-based surveys as the reference point are undermined by the fact that these may suffer from a high social-desirability bias. Studies using random samples are rare and there is a lack of information about unobserved sample members.

In this article, we approach the issues above by comparing a telephone survey with a web survey, both based on a probability sample. We use three designs with a randomized mode and incentive, all using the same questionnaire. The web survey includes a prepaid cash-like incentive experiment. We address sampling shortcomings in past research by using three samples that were drawn at random from an individual population register maintained by the SFSO. This register has the added advantage of containing the sociodemographic variables of age, sex, marital status, municipality size, and language region for all sample members.

We analyze completion rates and sample composition in the three designs (telephone, web without incentive, web with incentive). In addition, we compare voting behavior (turnout and party choice) with the actual election results in all three designs and finally discuss cost issues.

3. Data and Experimental Design

The Swiss quadrennial election survey Selects (http://www.selects.ch) set up a web experiment alongside the regular CATI survey in the context of the 2011 survey. The samples for all three designs were randomly drawn from the national individual register. The CATI sample was stratified by the 26 Swiss cantons (NUTS 3 level) with small cantons oversampled to a minimum of 100 respondents each (N = 8,162 adult citizens). The sample was matched against different telephone registers with a matching rate of 85%. The field period ran from October 24 (the day after the election) to November 25, 2011. No incentives were offered to CATI sample members. For the web survey, 1,481 additional Swiss citizens were selected in a simple random sample design from the national individual register and recruited offline.

All sample members received an advance letter with the university letterhead and signed by the director of the project. The basic content of all letters was the same: a description of the study, including its purpose and why it is important that the person responds, the length of the interview, which was estimated at around 30 minutes for both modes, and contact information in case of questions. The only difference between the letters in the two modes was that in one the modalities of the telephone interview were explained, whereas for the web sample, the individuals were asked to complete the survey online using the Selects 2011 URL (www.selects2011.ch) and a unique code. No special Internet equipment was provided in the context of the study. Web-sample members unable to access the Internet therefore had no possibility of taking the survey. Therefore, while the telephone-sample members were told that they would be called by an interviewer in the coming days, the web-sample members were expected to be proactive by accessing the questionnaire online.

In order to compensate for this difference, 485 of the 1,481 web-sample members received a 20 CHF (CHF = .82€ = 1.11$ (as of February 5 2014)) (prepaid) postal check with the advance letter, whereas 996 did not receive this incentive. Postal checks can be cashed at no cost at any post office in Switzerland. The wording of the letters for the two web-sample members was identical except for the additional paragraph explaining the incentive to those who received one. In both web designs, two reminders were sent to those who had not yet responded within an identical timeframe. The announcement letters were sent on Friday October 21 by regular mail so that they would be received on Monday October 24 or Tuesday October 25, and the reminder letters were sent by priority mail on Friday November 4 and Wednesday November 16, ensuring that they would be received on the following day. This was done to minimize the number of individuals completing the interview in the meantime and thus receiving a reminder even though they had already completed the survey. Standard mail was used for the advance letter because for logistical reasons it was not possible to send the letters from the university on a Saturday, and with priority mail many would already have received the letter on the Saturday before the elections. Priority mail was also the preferred method for conveying the reminders as some individuals might think a letter sent by standard mail was an advertisement. The final respondents included in the data responded on December 12.

The telephone and the web questionnaire were the same, with only slight changes made to the wording of the questions to adapt them to a written mode. Each telephone-sample member was called at different times and refusal conversions attempted using more experienced interviewers. We summarize the contact dates and materials sent for each mode in Table 1.

4. Completion Rates and Sample Composition

In this chapter, we analyze response rates (RR1; AAPOR 2011) in the three designs for the sample members distinguished by experimental design and the variables available from the sampling frame. Unlike in the telephone mode, a general problem of the web mode is that the two components of noncompletion (or nonobservation) – noncoverage and nonresponse – cannot be separated. For a discussion of undercoverage and nonresponse issues in the telephone survey used here, see Lipps et al. (2015). In a web survey, noncompletion can only be further analyzed in the case of a break off, in which the sample member was at least found and successfully contacted, but did not complete the survey. For clarity, we consider incomplete responses as nonresponses. Note that incomplete surveys play a minor role in the web designs since only 14 sample members (1.4%) broke off the questionnaire in the without-incentive design and nine sample members (1.9%) in the with-incentive design. For both telephone and web, people without access to the contact mode are deemed “noncompleters”. Total and group-specific completion rates (in proportions), as well as significance levels measured as chi^2-differences between completers and noncompleters across the three designs are presented in Table 2.

Telephone has only a slightly higher (6% points) overall completion rate than web without incentive. Individuals who are older than 56 years, married, women, and

Table 1. Contact dates and materials sent in Selects 2011

	Telephone (N = 8,162)	Web no incentive (N = 996)	Web with incentive (N = 485)
Field period	October 24 - November 25	October 24 - December 12	October 24 - December 12
Advance letter: Oct 21 (regular mail)	- Study description	- Study description with - Selects 2011 URL	- Study description with - Selects 2011 URL and - 20 CHF prepaid postal check
Reminder letter 1		November 4	November 4
Reminder letter 2 (both priority mail)		November 16	November 16

Table 2. Sample sizes and completion rates in the different designs. Data: Selects 2011

	Telephone		Web no incentive		Web with incentive	
Sampling frame variable category	N	Mean (standard error)	N	Mean (standard error)	N	Mean (standard error)
All	8,162	.29 (.004)	996	.23 (.01)*	485	.44 (.02)$^{*\#}$
18 to 30-year-olds	1,642	.26 (.01)	218	.24 (.03)	106	.51 (.05)$^{*\#}$
31 to 43-year-olds	1,612	.30 (.01)	218	.22 (.03)	99	.46 (.05)$^{*\#}$
44 to 56-year-olds	2,029	.31 (.01)	234	.30 (.03)	114	.50 (.05)$^{*\#}$
57 to 69-year-olds	1,594	.35 (.01)	199	.19 (.03)*	100	.46 (.05)$^{\#}$
70+ year-olds	1,285	.24 (.01)	127	.13 (.03)*	66	.17 (.05)
Not married	3,827	.25 (.01)	461	.22 (.02)	237	.39 (.03)$^{\#}$
Married	4,335	.33 (.01)	535	.23 (.02)*	248	.49 (.03)$^{*\#}$
>100K inhabitants	756	.27 (.02)	99	.18 (.04)	61	.39 (.06)$^{\#}$
20–100K inhabitants	872	.31 (.02)	113	.28 (.04)	64	.45 (.06)
10–20K inhabitants	1,356	.30 (.01)	183	.21 (.03)	74	.41 (.06)$^{\#}$
5–10K inhabitants	1,386	.29 (.01)	163	.21 (.03)	87	.51 (.05)$^{*\#}$
2–5K inhabitants	2,044	.28 (.01)	238	.24 (.03)	110	.48 (.05)$^{*\#}$
<2K inhabitants	1,748	.30 (.01)	200	.23 (.03)	89	.38 (.05)$^{\#}$
Women	4,222	.29 (.01)	507	.20 (.02)*	254	.46 (.03)$^{*\#}$
Men	3,940	.29 (.01)	489	.25 (.02)	231	.42 (.03)$^{*\#}$
Language Swiss German	6,128	.30 (.01)	739	.21 (.01)*	352	.46 (.03)$^{*\#}$
Language French or Italian	2,034	.28 (.01)	257	.28 (.03)	133	.39 (.04)*

Notes: Means design-weighted. $^{*,\#}$ means that the value is significantly different at the 1% level (Pearson chi^2 test) compared to the respective value in the: * telephone, $^{\#}$ web-without-incentive design.

Swiss-German speakers respond significantly (5% level) less in the web design. Comparing the two web designs, the incentive increases completion rates for all groups (21% points overall), and most comparisons are statistically significant except for the 70+ group.

In Table 3 we examine sample composition in the population, the total sample, and the three designs. Compared to census data, young people, those in smaller villages, and Swiss-German speakers are slightly overrepresented in the total sample. Comparing the total sample with the respondents in the three designs, individuals who are young, unmarried, from large municipalities (>100K inhabitants), and who speak French or Italian are underrepresented among the telephone respondents, and older individuals are underrepresented among the web respondents. The relative underrepresentation of older people becomes worse with the incentive. Although the composition of the web-with-incentive sample is quite different to that of the web-without-incentive sample, no group of persons is significantly different due to the small sample sizes. Because the chi^2-values are not comparable between the designs, we analyze sample bias of the three designs using the sum of the absolute percentage differences to the total sample across all 17 categories of the five sociodemographic variables. The maximum absolute percentage difference would be 500 percentage points if the samples were distributed completely differently across the five variables (e.g., in the case of sex, the sample in one design has 100% women, whereas the total sample has 100% men, and similarly for the other four variables). This method gives percentage point differences of 40.9 in the telephone design, 61.0 in the web-without-incentive design, and 49.8 in the web-with-incentive design across all categories. The telephone sample has the smallest representation bias for age, sex, and municipality size, the web-without-incentive sample has the smallest representation bias for marital status, and the web-with-incentive sample has the smallest representation bias for language group.

The group-specific completion differences and sample composition mostly reflect differential access to landline telephone or web (and its use), respectively (Alexander et al. 2008). On the one hand, compared with the average telephone matching rate (85%) of sample members from the individual register, the underrepresented person groups (see Table 3), the 18- to 30-year-olds (74%), the unmarried (79%), those in large municipalities (>100,000 inhabitants, 77%), and French (83%) or Italian (81%) speakers, are less likely to be matched. The 57 to 69-year-olds, who are overrepresented in the telephone sample, have a telephone matching rate of 92%. For people aged 70 or over, noncooperation is much higher than for other age groups, which is the reason why this age group is still underrepresented among the telephone respondents, in spite of its above-average matching rate (95%).

According to the SFSO's Omnibus (2010) survey, 78% of the population over 15 uses the Internet (at least once during the last three months), but the figure is much lower for older groups: 45% for those aged 65 to 74 and only 20% for those aged 75+. In our results, there is a reverse effect of telephone and Internet coverage: people with higher telephone coverage (older, Swiss-German speakers) tend to have lower Internet coverage and vice versa. Exceptions are married people who have both high telephone and high Internet coverage, probably due to economies of scale in larger households. For those with less web access (and lower computer literacy) such as older people, even an incentive is not able to substantially increase participation.

Table 3. Population and respondent sample distributions. Data: Selects 2011

				Web	
Sampling frame variable category	Census[a]	Total Sample	Telephone	No incentive	With incentive
18 to 30-year-olds	19.2	20.4	18.0+	23.1	25.2*
31 to 43-year-olds	20.0	20.0	20.2	21.8	21.5
44 to 56-year-olds	24.8	24.7	25.9	31.6	26.6
57 to 69-year-olds	19.9	19.6	23.4+	16.4	21.5
70+ year-olds	16.1	15.3	12.6+	7.1+	5.1+*
Not married	46.9	46.9	39.9+	45.8	43.0
Married	53.1	53.1	60.1+	54.2	57.0
>100K inhabitants	10.8	9.3	8.4	8.0	11.2
20–100K inhabitants	11.9	10.9	11.1	14.2	13.6
10–20K inhabitants	16.6	16.7	16.8	16.9	14.0
5–10K inhabitants	17.3	17.0	17.9	15.6	20.6
2–5K inhabitants	23.8	24.8	24.5	24.9	24.8
<2K inhabitants	19.7	21.2	21.3	20.4	15.9
Women	52.3	51.7	50.9	45.8	54.2
Men	47.8	48.3	49.1	54.2	45.8
Language Swiss German	74.0[b]	74.9	78.2+	68.0*	75.7
Language French or Italian	26.0[b]	25.1	21.8+	32.0	24.3
N	386,995	9,643	2,371	225	214

Notes: [a] Pooled Swiss structural surveys 2010 and 2011, person weighted. Differences between the census and the total sample could be due to sampling errors and the definition of second and third nationalities.

[b] (main) language region.

+ , *, $^{\#}$ means that the value is significantly different at the 1% level (Pearson chi^2 test) compared to the respective value in the: + total sample, * telephone, $^{\#}$ web no incentive design.

Comparisons of distributions (chi^2 test statistics not shown):

- Telephone distribution is significantly different (1%) from total sample distribution for age, municipality size, and language
- Web without incentive is significantly different (1%) from total sample distribution for age
- Web with incentive is significantly different (1%) from total sample distribution for age
- Web without incentive is significantly different (1%) from telephone distribution for age and language
- Web with incentive is significantly different (1%) from telephone distribution for age
- Web with incentive is significantly different (1%) from web without incentive for no demographic variable

5. Substantive Results

We analyze some key variables regarding political behavior, without and with sociodemographic control variables. The variables included in the models are: age group, gender, education, marital status, language region, municipality type, and religion. As already seen, the composition of the samples varies between designs. We wish to know whether the differences in substantive variables are due to this variation or whether there are other sources of variation, such as mode effects. We prefer sociodemographic control variables to weighting because the variables which the weights are based on can be included in the regression models easily, avoiding possible errors that can arise from creating weights (see e.g., Little and Vartivarian 2003). To control for sociodemographic variables, we use logistic regressions (Poisson regression for the two likert scale variables of political interest and participation) with the design as the sole independent variable in the first part of the table and in conjunction with the sociodemographic variables in the second. The predicted marginal values were calculated so as to compare coefficients across different models (Mood 2010).

Generally, we note that due to the small sample sizes in the web designs, the number of significant differences between the two web designs is very small, even though differences in absolute terms are large for many comparisons.

A particular strength of the Selects survey is that questions about political behavior are asked within a short timeframe immediately following the elections. This makes it possible to compare turnout and party choice with the official results, see Table 4. It is known that election surveys tend to overestimate turnout, due both to overreporting (social desirability) and selection bias (Burden 2000; Holbrook and Krosnick 2010; Karp and Brockington 2005; McDonald 2003; Selb and Munzert 2013). In our study, it appears that the web-with-incentive design is more accurate, where the result of 65.7% is eight percentage points closer to the actual figure of 48.5% compared to the telephone survey. Without incentive, the result of the web survey is very similar to that of the telephone survey. It would appear that in this case selection bias is probably a stronger reason for the differences than overreporting, as the main change occurs when an incentive is added and not between modes.

To analyze vote choice, a left/center/right variable was constructed. This combination is necessary due to the large number of parties in Switzerland and the ensuing low number of cases by party in the web conditions, especially after excluding non-voters. In Table 4, we find the distribution to be much closer to the actual election results for the web-with-incentive design. The change is particularly apparent in the case of right-wing parties, whose voters are strongly underrepresented in the two other conditions.

Individuals who are more interested and active in politics are generally easier to reach in surveys (see e.g., Groves et al. 2004a,b). In the web-without-incentive design, people appear to be more interested in politics and vote more often, while the opposite is true when an incentive is offered. The incentive thus seems essential if we wish to attenuate the bias towards those who are interested and active in politics in a web survey.

Even though the differences are not significant between the telephone and the web-with-incentive samples, there is a consistent tendency towards less politically interested individuals in the latter. We thus hypothesize that this design is also able to reach people

Table 4. Comparisons of substantive variables in CATI, web, and web-with-incentive samples, compared with official figures when available. Data: Selects 2011

					Includes sociodemographic controls		
	Official	Telephone	Web no inc.	Web w. inc.	Telephone	Web no inc.	Web w. inc.
Turnout	48.5	73.9 (.01)	73.2 (.03)	65.7 (.03)*	74.4 (.01)	75.1 (.03)	68.9 (.03)
N			2,807			2,668	
Block voted for							
Left	28.9	34.2 (.01)	33.1 (.04)	30.0 (.04)	34.4 (.01)	29.2 (.04)	30.2 (.04)
Center	25.7	27.8 (.01)	32.4 (.04)	23.1 (.04)	27.8 (.01)	33.6 (.04)	22.6 (.04)$^{\#}$
Right	45.4	37.9 (.01)	34.5 (.04)	46.9 (.04)$^{*\#}$	37.8 (.01)	37.7 (.04)	47.9 (.04)*
N			1,883			1,814	
Polit. Interest (1–4)		1.88 (.02)	1.95 (.05)	1.78 (.06)$^{\#}$	1.88 (.02)	1.97 (.05)	1.78 (.05)$^{\#}$
Participation in votes (0–10)		7.50 (.07)	8.00 (.19)*	7.28 (.21)$^{\#}$	7.50 (.07)	7.99 (.18)*	7.39 (.21)$^{\#}$
N			2,714			2,658	

Notes: Mean percentage for turnout and party block voted for (mean score for political interest and participation) and standard error in brackets.
$^{*,\#}$ means that the value is significantly different at the 5% level (Pearson chi^2 test) compared to the respective value in the: * telephone, # web no incentive design.

who would not respond to the traditional telephone survey. This is consistent with the inclusion of more right-wing voters. Our results show that the web-with-incentive design has certain advantages in obtaining a more representative sample of the studied population regarding political behavior.

We then turn to the results of the predicted probabilities for turnout and vote choice after controlling for sociodemographics in the right panel of Table 4. We find some changes compared to the uncontrolled figures. For instance, turnout figures become larger for the web design, in particular with incentive. This shows that some of the initial difference between web and telephone is due to the inclusion of sociodemographic groups who vote less in the web (with incentive). Regarding voting behavior, the numbers change slightly for the telephone. For the web modes, the tendency is that the difference between support for the right and the left increases in favor of the right, especially in the design without incentive. In addition, the difference in support for center parties becomes significant as the changes go in opposite directions. It would appear that while the web mode attracts fewer individuals with the sociodemographic characteristics of voters in general, and right-wing voters in particular (e.g., individuals from rural areas), the incentive is powerful in attracting right-wing voters. In turn, sociodemographic groups linked with voting for the left (e.g., individuals from urban areas) are overrepresented in the nonincentivized web survey. Some of the differences regarding voting behavior between the three conditions can thus be explained simply by the different sociodemographic composition of the net samples. However, we find this to be only a very small part of the explanation, and it does seem that the incentive, and to a lesser extent the mode, attracts individuals who are different regarding their political preferences and behavior, net of their sociodemographic profile. Finally, we tested primacy and recency effects in the two modes considered elsewhere (Lipps and Pekari 2013) and found them to be insignificant.

6. Cost Issues

As the final part of our comparison of modes, we briefly present the cost by mode. For the results to be more generalizable, we include two cost estimates for the web survey: the costs of the actual in-house survey and the cost if the survey had been done by a survey company. The costs are estimated based on two offers received during the call for tenders for the Selects 2011 project. The in-house web survey was carried out by the Selects team, consisting of a project leader, a junior researcher, a doctoral student, and students for the administrative tasks. The centralized CATI survey was outsourced to a survey company. The figures provide an indication of the cost savings that can be achieved by switching from telephone to web survey. In Table 5, we present a summary of the survey costs by design, assuming a targeted number of 2,000 respondents in each design. Costs for the incentivized version of the web are listed in brackets.

As can be expected, the web-without-incentive design is by far the cheapest mode, followed by web with incentive. Unit costs amount to about 23 CHF in the web-without-incentive survey, to 43 CHF in the web-with-incentive survey (both in house), and to 94 CHF in the telephone survey. If the web survey had been carried out by a survey company, the costs would have been 38 CHF without incentive, and 58 CHF with incentive. Interestingly, the incentive design cost only an additional 4 CHF per sample member,

Table 5. Calculated unit costs for the Selects 2011 survey (in CHF), assuming a targeted number of 2,000 respondents in each design. Figures in brackets are for the incentive condition. Data: Selects 2011

Cost Component	Telephone	Web In House	Web Survey Firm
CATI total (survey firm)	94.-		
Web total (survey firm)			38.- (30.-)
Programming and project management web		7.-	
Postage web (incl. reminders)		15.- (7.-)	
Coding open questions web		1.-	
Incentives web		(28.-#)	(28.-#)
Total	94.-	23.- (43.-)	38.- (58.-)

#Note: For a nominal value of 20 CHF, an uncashed postal check costs 3.5 CHF, a cashed postal check costs 23.5 CHF. 83% of the respondents cashed the check; of the nonrespondents, 13% cashed it. Given a response rate of 44%, the incentive per sample member costs .44*(.83*23.5 + .17*3.5) + .56*(.13*23.5 + .87*3.5) = 12.3 which makes 12.3/.44 = 28.0 per respondent.

although its nominal value amounts to 20 CHF, plus a 3.50 CHF administrative charge. Primarily, this is due to the fact that only a part of the sample members actually cash the check. Sending the same amount of cash would cost 20/.44 = 45.5 CHF per respondent. Ceteris paribus, the cost savings amount to 45.5–28.0 = 17.5. CHF. Secondly, as much as 8 CHF per sample member can be saved due to the almost twice as high response rate by sending much fewer letters (announcement and reminder) in the web-with-incentive condition. Whether this holds true for other surveys in other contexts needs further examination. Absolute cost figures will depend very much on the survey and cultural context, but we believe the cost structures to be roughly comparable in relative terms.

7. Summary of Findings

We find that *completion rates* in the telephone survey are only slightly higher than those in the web-without-incentive design, but considerably higher in the web-with-incentive design. Switching from telephone to web without incentive reduces completion rates for people who are 57 years and older, married, women, or Swiss-German speakers. The incentivized web survey increases participation almost across the board. The main concern with using the web design are those aged 70 and above, who are simply likely to be unable to complete the survey independent of their willingness to do so, whether for lack of an Internet connection or for the lack of the necessary computer literacy.

When calculated as the absolute percentage differences to the census, *sample composition* bias is highest in the web-without-incentive design, followed by the web-with-incentive design, and finally the telephone. While in both web designs older people are underrepresented, married people and Swiss-German speakers are overrepresented in the telephone design, and young people in both web designs, especially in the incentivized version. Swiss-German speakers are underrepresented in the web-without-incentive design.

As for *substantive outcomes*, the web-with-incentive design comes closest to the official voting results in terms of turnout and party choice. In addition, the mean values for political interest and voting participation, typically overestimated in election surveys, are lower than those for the other two designs. In contrast, telephone and web without incentive overestimate turnout and underestimate voting for right-wing parties. Sociodemographic controls lead to some changes, but do not explain the differences found between the different conditions.

Finally, the web survey *costs* much less, even with a 20 CHF incentive. A rough comparison between the three designs results in unit costs that are about four times as high in the telephone design compared to the web-without-incentive design (provided the web survey is conducted in house). Incentives increase the costs by about a sixth.

8. Conclusion

In the present study, we analyze the effects of two different modes (telephone and web) and a randomized incentive experiment within the Swiss Electoral Study (Selects) 2011 survey. The aim of this study is to analyze the effects of a possible switch from telephone to web due to cost and landline undercoverage issues. The innovation of our research is that we use a probability sample in each design. We focus on mode system effects (Biemer and Lyberg 2003). That is, we are interested in the systematic effects that are largely independent of (unavoidable) sampling and fieldwork-related differences. We examine completion rates and sample composition in the three designs distinguished by the frame variable characteristics. We then analyze substantive results and finally compare unit costs in the three designs.

The web-with-incentive design outperforms the web-without-incentive design not just in terms of response rates but also in regard to sample composition and substantive outcomes at comparatively small additional cost. It also outperforms the telephone mode on all accounts, except for a small disadvantage in terms of sample composition. In addition, the unit costs of the web-with-incentive survey are less than a half of that of the telephone survey. An incentivized web survey thus already appears to be highly competitive compared to the telephone survey in the context of an election study.

Compared to a survey of the general population, our findings must be seen as relating to the special case of an *election* survey. First, we only sample adults, and second, foreigners are excluded from the sample. Therefore, problems such as language, literacy, and – last but not least – motivation to complete the survey are likely to play a smaller role in our sample of adult Swiss citizens. Second, we are able to make use of an address register which allows individualized invitation letters to be sent to the sample members. Limitations include the comparatively small web-sample sizes, which make it difficult to compare distributions or sample statistics with the larger telephone survey. In addition, one would wish to have a full factorial design experiment, including a telephone-with-incentive design, to better evaluate the incentive effect in both modes. An important finding is that, while controlling for sociodemographics when comparing substantive results shows some sample selection tendencies, it is not able to explain selection effects related to other personal characteristics unrelated to sociodemographics (values, attitudes, etc.), which are likely to constitute the lion's share of the effects. Finally, we are unable to uncover whether substantive differences across the designs result from selection or mode (measurement) effects. We believe that fully identifying these effects is an important issue for further research (see, e.g., Vannieuwenhuyze and Loosveldt 2013). Nevertheless, we are confident that we have been able to demonstrate the strong potential of web surveys, compared with traditional telephone surveys, in an environment where technological and societal trends clearly speak in favor of this mode.

9. References

AAPOR (The American Association for Public Opinion Research). 2011. Standard definitions: Final dispositions of case codes and outcome rates for surveys, 7th ed. AAPOR.

Alexander, G., G. Divine, M. Couper, J. McClure, M. Stopponi, K. Fortman, D. Tolsma, V. Strecher, and C. Johnson. 2008. "Effect of Incentives and Mailing Features on Online Health Program Enrolment." *American Journal of Preventive Medicine* 34: 382–388. Doi: http://dx.doi.org/10.1016/j.amepre.2008.01.028.

Ansolabehere, S. and B. Schaffner. 2014. "Does Survey Mode Still Matter? Findings From a 2010 Multi-Mode Comparison." *Political Analysis* 22: 285–303. Doi: http://dx.doi.org/10.1093/pan/mpt025.

Atkeson, L., A. Adams, and R. Alvarez. 2014. "Nonresponse and Mode Effects in Self-and Interviewer-Administered Surveys." *Political Analysis* 22: 304–320. Doi: http://dx.doi.org/10.1093/pan/mpt049.

Baker, R., S.J. Blumberg, J.M. Brick, M.P. Couper, M. Courtright, J.M. Dennis, D. Dillman, M.R. Frankel, P. Garland, R.M. Groves, C. Kennedy, J. Krosnick, P.J. Lavrakas, S. Lee, M. Link, L. Piekarski, K. Rao, R.K. Thomas, and D. Zahs. 2010. "Research Synthesis AAPOR Report on Online Panels." *Public Opinion Quarterly* 74: 711–781. Doi: http://dx.doi.org/10.1093/poq/nfq048.

Biemer, P. 1988. "Measuring Data Quality." In *Telephone Survey Methodology*, edited by W. Nicholls II, R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II, and J. Waksberg, 273–283. New York: Wiley & Sons.

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

Birnholtz, J., D. Horn, T. Finholt, and S. Bae. 2004. "The Effects of Cash, Electronic, and Paper Gift Certificates as Respondent Incentives for a Web-Based Survey of Technologically Sophisticated Respondents." *Social Science Computer Review* 22: 355–362. Doi: http://dx.doi.org/10.1177/0894439304263147.

Blumberg, S. and J. Luke. 2013. *Wireless Substitution: Early Release of Estimates from the National Health Interview Survey*. July–December 2012. Available at: http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201306.pdf (accessed December 2014).

Bosnjak, M. 2005. "Effects of Two Innovative Techniques to Apply Incentives in Online Access Panels." Presentation at the General Online Research Conference (GOR), Zürich, March 22–23.

Bosnjak, M. and T. Tuten. 2003. "Prepaid and Promised Incentives in Web Surveys: an Experiment." *Social Science Computer Review* 21: 208–217. Doi: http://dx.doi.org/10.1177/0894439303021002006.

Brandtzæg, P., J. Heim, and A. Karahasanovic. 2011. "Understanding the New Digital Divide – A Typology of Internet Users in Europe." *International Journal of Human-Computer Studies* 69: 123–138. Doi: http://dx.doi.org/10.1016/j.ijhcs.2010.11.004.

Braunsberger, K., H. Wybenga, and R. Gates. 2007. "A Comparison of Reliability Between Telephone and Web-Based Surveys." *Journal of Business Research* 60: 758–764. Doi: http://dx.doi.org/10.1016/j.jbusres.2007.02.015.

Burden, B.C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8: 389–398.

Busse, B. and M. Fuchs. 2012. "The Components of Landline Telephone Survey Coverage Bias. The Relative Importance of No-Phone and Mobile-Only Populations." *Quality and Quantity* 46: 1209–1225. Doi: http://dx.doi.org/10.1007/s11135-011-9431-3.

Callegaro, M. and C. DiSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72: 1008–1032. Doi: http/dx.doi.org/10.1093/poq/nfn065.

Chang, L. and J. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet." *Public Opinion Quarterly* 73: 641–678. Doi: http://dx.doi.org/10.1086/346010.

Christian, L., D. Dillman, and J. Smith. 2008. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In *Advances in Telephone Surveys*, edited by J.M. Lepkowski, 250–275. New York: Wiley & Sons.

Cobben, F. and J. Bethlehem. 2005. "Adjusting Undercoverage and Nonresponse Bias in Telephone Surveys." Discussion paper 05006. CBS, Statistics Netherlands, Voorburg/Heerlen. Available at: http://www.cbs.nl/nr/rdonlyres/7fd00f42-15a3-4151-9daa-2d54-566cf59a/0/200506x10pub.pdf (accessed February, 2016).

Curtin, R., E. Singer, and S. Presser. 2007. "Incentives in Random Digit Dial Telephone Surveys: a Replication and Extension." *Journal of Official Statistics* 23: 91–105.

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

Dillman, D. 2000. *Mail and Telephone Surveys: The Tailored Design Method*. New York: John Wiley & Sons.

Dillman, D. 2011. *Mail and Internet surveys: The Tailored Design Method – 2007 Update With New Internet, Visual, and Mixed-Mode Guide*. New York: John Wiley & Sons.

Dillman, D. and L. 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.

Dillman, D., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B. Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research* 28: 1–18. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.00.

Ernst Stähli, M. 2012. "Telephone Surveys in Switzerland: Spotlight." In *Telephone Surveys in Europe: Research and Practice*, edited by M. Häder, S. Häder and M. Kühne, 25–36. Berlin: Springer.

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

Göritz, A. 2004. "The Impact of Material Incentives on Response Quantity, Response Quality, Sample Composition, Survey Outcome, and Cost in Online Access Panels." *International Journal of Market Research* 46: 327–345.

Göritz, A. 2006. “Incentives in Web Studies: Methodological Issues and a Review.” *International Journal of Internet Science* 1: 58–70.

Groves, R., R. Cialdini, and M. Couper. 1992. “Understanding the Decision to Participate in a Survey.” *Public Opinion Quarterly* 56: 475–493. Doi: http://dx.doi.org/10.1086/269338.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. 2004a. *Survey Methodology*, Wiley Series in Survey Methodology. New York: Wiley.

Groves, R., S. Presser, and S. Dipko. 2004b. “The Role of Topic Interest in Survey Participation Decisions.” *Public Opinion Quarterly* 68: 2–31. Doi: http://dx.doi.org/10.1093/poq/nfh002.

Holbrook, A.L. and J.A. Krosnick. 2010. “Social Desirability Bias in Voter Turnout Reports Tests Using the Item Count Technique.” *Public Opinion Quarterly* 74: 37–67. Doi: http://dx.doi.org/10.1093/poq/nfp065.

Joye, C. 2012. “Srph-Castem.” FORS – SFSO workshop, June 21. Neuchâtel.

Joye, D., A. Pollien, M. Sapin, and M. Ernst Stähli. 2012. “Who Can Be Contacted by Phone? Lessons from Switzerland.” In *Telephone Surveys in Europe: Research and Practice*, edited by M. Häder, S. Häder and M. Kühne, 85–102. Berlin: Springer-Verlag.

Karp, J.A. and D. Brockington. 2005. “Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries.” *The Journal of Politics* 67: 825–840. Doi: http://dx.doi.org/10.1111/j.1468-2508.2005.00341.x.

Kreuter, F., S. Presser, and R. Tourangeau. 2008. “Social Desirability Bias in CATI, IVR, and Web Surveys.” *Public Opinion Quarterly* 72: 847–865. Doi: http://dx.doi.org/10.1093/poq/nfn063.

Krosnick, J. 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.

Link, M. and M. Fahimi. 2008. “Telephone Survey Sampling.” In *Sampling of Populations: Methods and Applications*, edited by P.S. Levy and S. Lemeshow, 455–487. New York: Wiley.

Lipps, O. and K. Kissau. 2012. “Nonresponse in an Individual Register Sample Telephone Survey in Lucerne (Switzerland).” In *Telephone Surveys in Europe: Research and Practice*, edited by M. Häder, S. Häder and M. Kühne, 187–208. Berlin: Springer-Verlag.

Lipps, O. and N. Pekari. 2013. *Mode and Incentive Effects in an Individual Register Frame Based Swiss Election Study*. FORS Working Paper Series, paper 2013-3. Lausanne: FORS.

Lipps, O., N. Pekari, and C. Roberts. 2015. “Coverage and Nonresponse Errors in an Individual Register Frame Based Swiss Telephone Election Study.” *Survey Research Methods* 9: 71–82.

Little, R.J. and S. Vartivarian. 2003. “On Weighting the Rates in Non-Response Weights.” *Statistics in Medicine* 22: 1589–1599. Doi: http://dx.doi.org/10.1002/sim.1513.

Lozar Manfreda, K., M. Bosnjak, J. Berzelak, I. Haas, and V. Vehovar. 2008. “Web Surveys Versus Other Survey Modes: a Meta-Analysis Comparing Response Rates.” *International Journal of Market Research* 50: 79–104.

Malhotra, N. and J. Krosnick. 2007. “The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys With Nonprobability Samples.” *Political Analysis* 15: 286–323. Doi: http://dx.doi.org/10.1093/pan/mpm003.

McDonald, M.P. 2003. “On the Overreport Bias of the National Election Study Turnout Rate.” *Political Analysis* 11: 180–186. Doi: http://dx.doi.org/10.1093/pan/mpg006.

Messer, B.L. and D.A. Dillman. 2011. “Surveying the General Public Over the Internet Using Address-Based Sampling and Mail Contact Procedures.” *Public Opinion Quarterly* 75: 429–457. Doi: http://dx.doi.org/10.1093/poq/nfr021.

Mohorko, A., E. de Leeuw, and J. Hox. 2013a. “Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time.” *Journal of Official Statistics* 29: 609–622. Doi: http://dx.doi.org/10.2478/jos-2013-0042.

Mohorko, A., E. de Leeuw, and J. Hox. 2013b. “Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage Across Countries and Over Time.” *Survey Methods: Insights from the Field*. Doi: http://dx.doi.org/10.13094/SMIF-2013-00002.

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.

Nagelhout, G., M. Willemsen, M. Thompson, G. Fong, B. van den Putte, and H. de Vries. 2010. “Is Web Interviewing a Good Alternative to Telephone Interviewing? Findings from the International Tobacco Control (ITC) Netherlands Survey.” *BMC Public Health* 10: 351. Doi: http://dx.doi.org/10.1186/1471-2458-10-351.

Omnibus 2010. *Survey on Information and Communication Technology, Swiss Federal Statistical Office 2010*. Excel result sheets (in German; accessed October 21, 2013). Available at: http://www.bfs.admin.ch/bfs/portal/de/index/themen/16/04/data.html (accessed December 2014).

Parsons, N. and M. Manierre. 2014. “Investigating the Relationship Among Prepaid Token Incentives, Response Rates, and Nonresponse Bias in a Web Survey.” *Field Methods* 26: 191–204. Doi: http://dx.doi.org/10.1177/1525822X13500120.

Peytchev, A. 2009. “Survey Breakoff.” *Public Opinion Quarterly* 73: 74–97. Doi: http://dx.doi.org/10.1093/poq/nfp014.

Revilla, M.A. and W.E. Saris. 2013. “A Comparison of the Quality of Questions in a Face-to-Face and a Web Survey.” *International Journal of Public Opinion Research* 25: 242–253. Doi: http://dx.doi.org/10.1093/ijpor/eds007.

Ryu, E., M. Couper, and R. Marans. 2006. “Survey Incentives: Cash vs. In-Kind, Face-to-Face vs. Mail, Response Rate vs. Nonresponse Error.” *International Journal of Public Opinion Research* 18: 89–106. Doi: http://dx.doi.org/10.1093/ijpor/edh089.

Sánchez-Fernández, J., F. Muñoz-Leiva, F.J. Montoro-Ríos, and J. Ángel Ibáñez-Zapata. 2010. “An Analysis of the Effect of Pre-Incentives and Post-Incentives Based on Draws on Response to Web Surveys.” *Quality and Quantity* 44: 357–373. Doi: http://dx.doi.org/10.1007/s11135-008-9197-4.

Schaurer, I., B. Struminskaya, L. Kaczmirek, and W. Bandilla. 2012. “The Price We Have to Pay: Incentive Experiments in the Recruitment Process for a Probability-Based

Online Panel." Presentation at the General Online Research Conference (GOR) March 5–7, 2012, Mannheim.

Scherpenzeel, A. and V. Toepoel. 2012. "Recruiting a Probability Sample for an Online Panel. Effects of Contact Mode, Incentives, and Information." *Public Opinion Quarterly* 76: 470–490. Doi: http://dx.doi.org/10.1093/poq/nfs037.

Schonlau, M., A. van Soest, A. Kapteyn, and M. Couper. 2009. "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods and Research* 37: 291–318. Doi: http://dx.doi.org/10.1177/0049124108327128.

Selb, P. and S. Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 32: 186–196. Doi: http://dx.doi.org/10.1016/j.electstud.2012.11.004.

Sinclair, M., J. O'Toole, M. Malawaraarachchi, and K. Leder. 2012. "Comparison of Response Rates and Cost-Effectiveness for a Community-Based Survey: Postal, Internet and Telephone Modes with Generic or Personalised Recruitment Approaches." *BMC Medical Research Methodology* 12: 132. Doi: http://dx.doi.org/10.1186/1471-2288-12-132.

Singer, E. and R. Bossarte. 2006. "Incentives for Survey Participation. When are they 'Coercive'?" *American Journal of Preventive Medicine* 31: 411–418. Doi: http://dx.doi.org/10.1016/j.amepre.2006.07.013.

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

Singer, E. and C. Ye. 2013. "The Use and Effects of Incentives in Surveys." *The ANNALS of the American Academy of Political and Social Science* 645: 112–141. Doi: http://dx.doi.org/10.1177/0002716212458082.

Stephenson, L. and J. Crête. 2011. "Studying Political Behavior: A Comparison of Internet and Telephone Surveys." *International Journal of Public Opinion Research* 23: 24–55. Doi: http://dx.doi.org/10.1093/ijpor/edq025.

Struminskaya, B., L. Kaczmirek, I. Schaurer, and W. Bandilla. 2014. "Assessing Representativeness of a German Probability-Based Panel." In *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. Krosnick, and P. Lavrakas, 61–85. New York: John Wiley & Sons.

Su, J., P. Shao, and J. Fang. 2008. "Effect of Incentives on Web-Based Surveys." *Tsinghua Science and Technology* 13: 344–347. Doi: http://dx.doi.org/10.1016/S1007-0214(08)70055-5.

Teisl, M., B. Roe, and M. Vayda. 2006. "Incentive Effects on Response Rates, Data Quality, and Survey Administration Costs." *International Journal of Public Opinion Research* 18: 364–373. Doi: http://dx.doi.org/10.1093/ijpor/edh106.

Vannieuwenhuyze, J. and G. Loosveldt. 2013. "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects." *Sociological Methods & Research* 42: 82–104. Doi: http://dx.doi.org/10.1177/0049124112464868.

Van Veen, F., A. Göritz, and S. Sattler. 2011. "The Impact of Monetary Incentives on Completion and Data Quality in Online Surveys." Presentation at the European Survey

Research Association (ESRA) Conference, Lausanne, July 18–22 and General Online Research (GOR) Conference, Düsseldorf, March 14–16.

Von der Lippe, E., P. Schmich, and C. Lange. 2011. "Advance Letters as a Way of Reducing Non-Response in a National Health Telephone Survey: Differences Between Listed and Unlisted Numbers." *Survey Research Methods* 5: 103–116. Doi: http://dx.doi.org/10.18148/srm/2011.v5i3.4657#sthash.qLueRYqS.dpuf.

Warren, J. and A. Halpern-Manners. 2012. "Panel Conditioning in Longitudinal Social Science Surveys." *Sociological Methods and Research* 41: 491–534. Doi: http://dx.doi.org/10.1177/0049124112460374.

Yeager, D., J. Krosnick, L. Chang, H. Javitz, M. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75: 709–747. Doi: http://dx.doi.org/10.1093/poq/nfr020.

Zickhur, K. and A. Smith. 2012. *Digital Differences. Pew Internet & American Life Project* 13. Available at: http://www.pewinternet.org/2012/04/13/digital-differences/ (accessed August 2014).

Received February 2014
Revised March 2015
Accepted April 2015

DE
G
DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 187–208, http://dx.doi.org/10.1515/JOS-2016-0009

Bayesian Predictive Inference of a Proportion Under a Twofold Small-Area Model

Balgobin Nandram[1]

We extend the twofold small-area model of Stukel and Rao (1997; 1999) to accommodate binary data. An example is the Third International Mathematics and Science Study (TIMSS), in which pass-fail data for mathematics of students from US schools (clusters) are available at the third grade by regions and communities (small areas). We compare the finite population proportions of these small areas. We present a hierarchical Bayesian model in which the first-stage binary responses have independent Bernoulli distributions, and each subsequent stage is modeled using a beta distribution, which is parameterized by its mean and a correlation coefficient. This twofold small-area model has an intracluster correlation at the first stage and an intercluster correlation at the second stage. The final-stage mean and all correlations are assumed to be noninformative independent random variables. We show how to infer the finite population proportion of each area. We have applied our models to synthetic TIMSS data to show that the twofold model is preferred over a onefold small-area model that ignores the clustering within areas. We further compare these models using a simulation study, which shows that the intracluster correlation is particularly important.

Key words: Intracluster and intercluster correlations; credible intervals; goodness of fit; hierarchical model; simulation study.

1. Introduction

We assume that there are several small areas and each area consists of several clusters; each cluster consists of a number of units (individuals). A random sample of clusters is taken from each area and within each sampled cluster a random sample of units is taken. This is the twofold sample design. A hierarchical Bayesian model is used to make inference about the finite population proportion of each small-area. In this model we have an intracluster (between two units in the same cluster) correlation at the first stage and an intercluster (between two units in two different clusters in the same area) correlation at the second stage. We show that the intracluster correlation is important by comparing the

[1] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, U.S.A. Email: balnan@wpi.edu

Acknowledgments: Part of this work was first presented by Balgobin Nandram in honor of Jon Rao on his Seventy-fifth Birthday Conference at Carleton University, Canada. In addition, presentations were made at several universities and conferences worldwide. This work was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). The author is grateful to the three reviewers for their careful reading of the manuscript and their extensive reports. He is also grateful to the Co-Editor-in-Chief, Dr. Ingegerd Jansson, for encouragement. Further research on this topic is currently being pursued by my PhD students.

Dedication: This article is dedicated to my late parents, Chandramattie Boodram and Nandram Boodram, who have enormously influenced my thinking. Both parents passed away while I was doing this work.

twofold small-area model with a onefold small-area model (the intracluster correlation is ignored). The Third International Mathematics and Science Study (TIMSS) uses a similar design.

In Subsection 1.1 we describe the TIMSS data that we use to illustrate our methodology and we discuss its importance. In Subsection 1.2 we introduce pertinent literature to show what has been done in twofold modeling and related problems. In Subsection 1.3 we clearly identify the innovations in this paper. Finally, we show a plan of the entire article.

1.1. Description of TIMSS Data

TIMSS is sponsored by the International Association for the Evaluation of Education Achievement, an international organization of national research institutions and government research agencies, and it is used to compare the performance of primary school students in mathematics and science. TIMSS provides reliable and timely data on the mathematics and science achievement of third-grade US students compared to that of students in other countries. Of course, there are other studies used for this purpose with similar objectives (e.g., the Program for International Student Assessment, PISA). These studies provide information to “No Child Left Behind” and the “Race to the Top” programs in the US; to date, the US has spent more than ten billion dollars on the Race to the Top program since it was announced by President Barack Obama in 2009 (Hamilton 2009). Our study can potentially be used to suggest which regions and communities in the US need funding to improve the education systems (e.g., qualified teachers, improved equipment, parental participation, extramural programs, etc).

The basic sample design used in TIMSS for the population of third and fourth grade students was a two-stage stratified cluster design. The first stage consisted of a sample of schools; the second stage consisted of samples of one mathematics classroom from each eligible target grade in the sampled schools. The design required schools to be sampled using a probability proportional to size (PPS) systematic sampling (Foy et al. 1996), and classrooms to be sampled with equal probabilities. Different aspects of the design were adapted to national conditions and analytical needs. For example, many countries stratified the school sampling frame by variables of national interest. As another example, if geographic regions were an explicit stratification variable, then separate school sampling frames would be constructed for each region. The multistage stratified cluster design results in differential probabilities of selection and each student consequently has different weights. In a realistic analysis of the TIMSS data we would need to incorporate the survey weights into the analysis. However, because our main interest is to show how to handle the clustering within small areas, we have ignored the survey weights.

The data set, which we used and collected in 1999, consists of 2,477 students (135 schools) who participated in TIMSS (see Calsyn et al. 1999). Clusters are schools while the units within the clusters are the students. Areas are formed crossing region and community. There are four regions of the US (Northeast, South, Central, and West) and there are three communities (village or rural area, outskirts of a town or city, and close to the center of a town or city), which the students come from. Thus there are twelve areas (strata). The binary variable is whether a student’s mathematics score is below average. We use synthetic data to illustrate our methodology and we take roughly half of the sampled data (i.e., a simple random sample of half the number of schools and a simple random sample of half the number of third-grade students from each selected school) for analysis and we use the other half to assess the predictive power of our procedure. The finite population is the original sample. Our objective is to make inference about the finite population proportion of students who earned below average scores in mathematics for each small-area. This measure can be used to compare the regions and the communities in the US.

The data (half) on the mathematics test scores are shown in Table 1, where we define the twelve areas (e.g., NR is a village or rural area in the north east). There are some schools in which all students were either below average or above average, thereby creating some difficulties for estimation. Looking at Table 1, the numbers of schools sampled in the twelve areas are 2, 4, 5, 4, 8, 6, 1, 3, 7, 3, 6, 15 and the numbers of students sampled in the schools range from 4 to 13. Each area is too sparse for direct estimation even with the complete data set.

1.2. Pertinent Literature

Nandram and Sedransk (1993) described a hierarchical Bayesian model to make inference about the finite population proportion under two-stage cluster sampling, the design we have within each area in a twofold sample design. The model can be viewed as a discrete analogue of the model for two-stage cluster sampling with normal data (Scott and Smith 1969) that has been extended in many directions (e.g., Malec and Sedransk 1985). We note that the work of Nandram and Sedransk (1993) was extended by Nandram (1998) to multinomial data and this extension may be viewed as a Bayesian analogue of the Dirichlet-multinomial model for cluster sampling (Brier 1980). However, our onefold model is different because in this design a simple random sample is taken from each area, but in the twofold model a two-stage cluster sample is performed in each area.

When there is a clustering effect, the units in a cluster are, in general, positively correlated leading to a smaller effective sample size and therefore larger variability in the estimates of the cell probabilities (i.e., the design effect is larger than one for each area). For example, see Brier (1980), Bedrick (1983), Holt et al. (1980), and Scott and Holt (1982). There is a similar issue in hypothesis testing. Clustering will evidently result in larger p-values than what would be obtained under simple random sampling. Rao and Scott (1981; 1984) have studied this problem very carefully for contingency tables and obtained simple and familiar corrections to the standard chi-squared statistic for the test of independence for two-way contingency tables arising from two-stage cluster sampling and more generally. Nandram et al. (2013) have a Bayesian analogue of these works.

From a Bayesian perspective, a related problem is when data are fitted to a hierarchical model but actually follow a model with an additional unknown structure. This is like our problem in which a onefold model is fitted and the second-stage cluster sampling within each area is ignored. Using posterior predictive p-values, Yan and Sedransk (2007) studied the situation where the data follow a normal model with a two-stage (three-stage) hierarchical structure while the fitted model has a one-stage (two-stage) hierarchical structure.

Table 1. Number of US students below average in mathematics within schools by area (region by community)

Area			*m*	Schools														
NR	s	9	2	4	5													
	n	18		8	10													
NO	s	12	4	5	5	1	1											
	n	38		10	12	7	9											
NC	s	39	5	9	11	6	13	0										
	n	56		11	12	12	13	8										
SR	s	24	4	7	4	9	4											
	n	37		10	7	11	9											
SO	s	31	8	3	2	0	8	7	2	9	0							
	n	81		9	10	8	9	13	9	12	11							
SC	s	43	6	8	6	4	7	8	10									
	n	52		8	8	9	8	9	10									
CR	s	4	1	4														
	n	11		11														
CO	s	13	3	7	4	2												
	n	24		11	9	4												
CC	s	41	7	10	6	7	6	3	5	4								
	n	60		10	6	9	10	8	9	8								
WR	s	8	3	1	4	3												
	n	27		8	8	11												
WO	s	27	6	3	6	8	5	3	2									
	n	46		6	6	10	8	8	8									
WC	s	83	15	4	5	7	10	10	3	6	6	2	8	6	6	2	5	3
	n	118		9	5	7	11	11	5	13	6	8	9	7	8	5	7	7

NOTE: For each area the counts are in pairs (*s*,*n*) with *s* at top and *n* bottom. The areas are formed by crossing region (N: north, S: south, C: central, W: west) and community (R: rural, O: outskirt of a town or city, C: town or city). For example, in area NR there are $m = 2$ schools with a total of $n = 18$ students (8 from one school and 10 from the other) and $s = 9$ students scored below average (4 from one school and 5 from the other). This constitutes approximately half the number of schools and half the number of students.

They used several diagnostic procedures to help detect this additional structure. Yan and Sedransk (2010) studied the ability to detect a three-stage model when a two-stage model is actually fitted, and using Bayesian standardized residuals concluded that it is due to the magnitude of correlation induced by the additional structure. This is the key point of our work.

For twofold modeling, there have been some activities for continuous response variables, not binary response variables, and most of this work has been within the empirical Bayes framework. Onefold and twofold nested error regression models were introduced by Fuller and Battese (1973) in which transformations to uncorrelated errors with constant variance are obtained starting with a general error covariance matrix. Transformations permit the calculation of generalized least-squares estimators and their covariance matrices by ordinary least-squares regression. They have made an analogy between survey sampling and experimental design via subsampling of primary, secondary, and tertiary sampling units, and split-split-plot experiments. Ghosh and Lahiri (1988) studied multistage sampling under posterior linearity using Bayes and empirical Bayes methods. Estimation of regression models with nested error structure and unequal error variances were further studied by Stukel and Rao (1997) under two-stage and three-stage cluster sampling. Small-area models under twofold nested error regression models were also studied by Stukel and Rao (1999); see also Rao (2003, sec. 5.5.3) and Datta and Ghosh (1991).

1.3. *Innovations*

This is mainly a methodological article on twofold small-area modeling, and in attempting to analyze the TIMSS data, we have made the following significant innovations.

1. Our models are for categorical data (binary). As can be seen from the literature, twofold modeling has been done for continuous data. While the categorical data models are related to the continuous data models, they pose additional difficulties for methodology and model fitting.
2. We have a new reparameterization of the beta distribution in terms of correlation (intracluster and intercluster). This permits modeling these correlations directly. In fact, this opens up a new avenue for the analysis of data collected using a twofold sample design and further analysis of more complex categorical (e.g., polychotomous) data.
3. With these reparameterizations we develop two hierarchical Bayesian models, a onefold and a twofold model for binary data.
4. The computations pose some difficulties for the Gibbs sampler and we have overcome these difficulties using random samples instead of the Gibbs sampler. In our twofold model there are two weakly identified parameters, thereby causing long-range dependence in the Gibbs sampler.
5. The TIMSS data will be analyzed using both our onefold and twofold models. We demonstrate that the intracluster correlation creates an important difference between the two models and provides additional insight to the analysis of these data. A simulation study demonstrates the importance of the twofold model for TIMSS data as well.

In Section 2 we describe the onefold and twofold models and we describe how to fit them. In Appendix A we describe how to perform the computation for the twofold model without using the Gibbs sampler. In a technical report, Nandram (2014) now called TRN14, we compare our sampling-based method with the Gibbs sampler. In Section 3, we analyze the TIMSS data and we also compare the onefold and twofold models. We also present a simulation study to compare the onefold and the twofold small-area model even further. Section 4 contains concluding remarks, and some additional problems are discussed. In Appendix B we briefly describe a multifold model.

2. Bayesian Small-Area Models

We make two simple observations. Let $y_i|p \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \sim \text{Beta}\{\mu\tau, (1-\mu)\tau\}$, where $0 < \mu < 1$ is the mean of the beta random variable and τ is the sum of the parameters of the standard beta distribution.

First, the y_i are exchangeable and the correlation between y_i and y_j is $\rho = (1 + \tau)^{-1}$ with $\tau = (1-\rho)/\rho$. Thus, we can write the model as $y_i|p \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \sim \text{Beta}\left\{\mu\frac{1-\rho}{\rho}, (1-\mu)\frac{1-\rho}{\rho}\right\}$.

Second, considering a single observation, y_1 say, the posterior mean of p given y_1 is

$$\text{E}(p|\rho, \mu, y_1) = \rho y_1 + (1-\rho)\mu.$$

The prior density, $\rho \sim \text{Uniform}(0,1)$, is called a shrinkage prior. Shrinkage priors have good frequentist properties (see Natarajan and Kass 2000; Molina et al. 2014; Toto and Nandram 2010). These observations motivate the construction of our small-area model for binary data.

We have a population of ℓ small areas and within the i^{th} area there are M_i clusters. Within the j^{th} cluster there are N_{ij} individuals. The binary responses are $y_{ijk}, k = 1, \ldots, N_{ij}$, $j = 1, \ldots, M_i$, $i = 1, \ldots, \ell$. A simple random sample of m_i clusters is taken from the i^{th} area and a simple random sample of n_{ij} individuals is taken from the j^{th} cluster. Let $n_i = \sum_{j=1}^{m_i} n_{ij}$, $s_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$, $s_i = \sum_{j=1}^{m_i} s_{ij}$.

Letting $N_i = \sum_{j=1}^{M_i} N_{ij}$, the finite population proportion for the i^{th} area is

$$P_i = \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} y_{ijk}/N_i, i = 1, \ldots, \ell.$$

Let $T_{ij}^{(1)} = \sum_{k=n_{ij}+1}^{N_{ij}} y_{ijk}$, $j = 1, \ldots, m_i$, denote the nonsampled total of the j^{th} sampled clusters and $T_{ij}^{(2)} = \sum_{k=1}^{N_{ij}} y_{ijk}, j = m_i + 1, \ldots, M_i$, the total of the j^{th} nonsampled cluster. Letting $n_i = \sum_{j=1}^{m_i} n_{ij}$, $\hat{p}_i = \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} y_{ijk}/n_i$, it is convenient to express P_i as

$$P_i = \left\{ n_i\hat{p}_i + \sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)} \right\} / N_i, i = 1, \ldots, \ell, \tag{1}$$

where the $\hat{p}_i$ are observed. Bayesian predictive inference is required for $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$. There is an expression similar to (1) for the finite population mean for each area (Stukel and Rao 1999).

2.1. A Onefold Model

We first construct the small-area onefold Bayesian model,

$$y_{ijk}|p_i \overset{ind}{\sim} \text{Bernoulli}(p_i), j = 1, \ldots, M_i, k = 1, \ldots, N_{ij},$$

$$p_i|\theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{ \theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma} \right\}, i = 1, \ldots, \ell,$$

where in a standard Beta(α, β), $\theta = \alpha/(\alpha + \beta)$ and $\gamma = (\alpha + \beta + 1)^{-1}$. Note that the cluster effects are dropped (i.e., the p_i do not have subscript j). Noting that θ and γ are really probabilities, a priori

$$\theta, \gamma \overset{iid}{\sim} \text{Beta}(\alpha_o, \beta_o),$$

where $\alpha_o = \beta_o$ for a noninformative prior with small values (e.g., $\alpha_o = 1$ for a uniform prior and $\alpha_o = .5$ for Jeffreys prior). Here, $0 < \gamma < 1$ strictly, and the uniform prior on γ is a shrinkage prior.

The model of Nandram and Sedransk (1993) for two-stage cluster sampling with binary responses is similar to the current one. One important difference is in the prior specification of θ and the reparametrization of γ, which unlike Nandram and Sedransk (1993) is stochastic here. Furthermore, we predict the finite population proportion of each area, not the overall finite population proportion.

The onefold model can be fitted easily by making random draws from the joint posterior density of θ and γ, and samples of p_i can be obtained using the multiplication rule. Specifically,

$$p_i|s_i, \theta, \gamma \overset{ind}{\sim} \text{Beta}\left\{ s_i + \theta\frac{1-\gamma}{\gamma}, n_i - s_i + (1-\theta)\frac{1-\gamma}{\gamma} \right\},$$

and

$$\pi(\theta, \gamma|\underset{\sim}{y}) \propto \prod_{i=1}^{\ell} \frac{B\{s_i + \theta(1-\gamma)/\gamma, n_i - s_i + (1-\theta)(1-\gamma)/\gamma\}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}} \times \theta^{\alpha_o - 1}$$

$$(1-\theta)^{\beta_o - 1}\gamma^{\alpha_o - 1}(1-\gamma)^{\beta_o - 1}, 0 < \theta, \gamma < 1, \tag{2}$$

where $B(\cdot\,,\cdot)$ is the beta function.

Because the posterior density of (θ,γ) is not in a simple form, we use a one-dimensional grid method and numerical integration via Gaussian quadrature to draw samples from it. We first integrate out θ to get $\pi(\gamma|\underset{\sim}{y}) \approx \sum_{g=1}^{G} w_g \pi(x_g, \gamma|\underset{\sim}{y})$, where $x_g, g = 1, \ldots, G$, are the G roots of a Legendre orthogonal polynomial with weights $w_g, g = 1, \ldots, G$; $G = 20$ or so provides a very accurate and fast procedure. Then, we use a one-dimensional grid to draw γ from $\pi(\gamma|\underset{\sim}{y})$. The unit interval is simply divided into 100 subintervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the midpoints of these subintervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\gamma|\underset{\sim}{y})$. It is efficient to remove subintervals with small probabilities (smaller than 10^{-6}); we call the others probable subintervals. To draw a single deviate, we first draw one of the probable subintervals. After we have obtained this subinterval, a uniform random variable is drawn within this subinterval. This is a standard jittering procedure and it provides different deviates with probability one. We call this random number generator the univariate grid sampler that is also used to fit the twofold model.

Once samples of the p_i are obtained, Bayesian predictive inference follows easily because $T_{ij}^{(1)}|p_i \overset{ind}{\sim} \text{Binomial}(N_{ij} - n_{ij}, p_i)$ and $T_{ij}^{(2)}|p_i \overset{ind}{\sim} \text{Binomial}(N_{ij}, p_i)$ and, given p_i, $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ are independent. It follows easily that $\sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}|p_i \sim \text{Binomial}(N_i - n_i, p_i)$. Thus it is easy to make inference about P_i by using data augmentation. For each iterate p_i, we simply draw $\sum_{j=1}^{m_i} T_{ij}^{(1)} + \sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}$. We use 1,000 samples; convergence monitoring is not required.

2.2. *A Twofold Model*

The twofold small-area model adds one layer to the onefold model. For a twofold Bayesian model,

$$y_{ijk}|p_{ij} \overset{ind}{\sim} \text{Bernoulli}(p_{ij}), k = 1, \ldots, N_{ij},$$

$$p_{ij}|\mu_i, \rho \overset{ind}{\sim} \text{Beta}\left\{\mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho}\right\}, j = 1, \ldots, M_i,$$

$$\mu_i|\theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{\theta \frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right\}, i = 1, \ldots, \ell,$$

and a priori

$$\rho, \theta, \gamma \overset{iid}{\sim} \text{Beta}(\alpha_o, \beta_o)$$

with the same comments about this prior as for the onefold model. We assume that $0 < \theta, \rho, \gamma < 1$ strictly. This can be achieved by taking $\epsilon \leq \theta, \rho, \gamma \leq 1 - \epsilon$, where ϵ is a small positive quantity (e.g., $\epsilon = 10^{-6}$).

If we allow ρ to go to zero, then the p_{ij} almost surely go to the μ_i and the twofold model becomes the onefold model. (In the limit, the μ_i in the twofold model become the p_i in the onefold model.) That is, if ρ is small, we anticipate very little difference between the two models. Thus it is ρ that distinguishes the onefold and twofold models.

In Subsection 2.1 we stated that $\text{cor}(y_{ijk}, y_{ijk'}|\mu_i, \rho) = \rho, k \neq k'$. That is, within the same area, the correlation between two units in the same cluster (intracluster) is ρ. Clearly, $\text{cor}(y_{ijk}, y_{ij'k'}|\mu_i, \rho) = 0$ and within the same area the actual correlation between two units in two different clusters (intercluster) is 0. It is also easy to show that

$$\text{cor}(y_{ijk}, y_{ij'k'}|\theta, \rho, \gamma) = \gamma, j \neq j', k \neq k'.$$

That is, one can interpret γ as the intercluster correlation between two units in two different clusters in the same area. Finally, note that $\text{cor}(y_{ijk}, y_{ijk'}|\theta, \rho, \gamma) = \gamma + (1-\gamma)\rho \geq \max(\rho, \gamma)$.

Using Bayes' theorem and letting $s_{ij}=\sum_{k=1}^{n_{ij}} y_{ijk}$, $\underset{\sim}{p}=(p_{ij}, j=1,\ldots,m_i, i=1,\ldots,\ell)'$, and $\underset{\sim}{\mu}=(\mu_i, i=1,\ldots,\ell)'$, the joint posterior density is

$$\pi(\underset{\sim}{p},\underset{\sim}{\mu},\underset{\sim}{\theta},\rho,\gamma|\underset{\sim}{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i} p_{ij}^{s_{ij}}(1-p_{ij})^{n_{ij}-s_{ij}}\frac{p_{ij}^{\mu_i(1-\rho)/\rho-1}(1-p_{ij})^{(1-\mu_i)(1-\rho)/\rho-1}}{B\{\mu_i(1-\rho)/\rho,(1-\mu_i)(1-\rho)/\rho\}}$$

$$\times\left\{\prod_{i=1}^{\ell}\frac{\mu_i^{\theta(1-\gamma)/\gamma-1}(1-\mu_i)^{(1-\theta)(1-\gamma)/\gamma-1}}{B\{\theta(1-\gamma)/\gamma,(1-\theta)(1-\gamma)/\gamma\}}\right\}\theta^{\alpha_o-1}(1-\theta)^{\beta_o-1}\rho^{\alpha_o-1}(1-\rho)^{\beta_o-1}\gamma^{\alpha_o-1}$$

$$(1-\gamma)^{\beta_o-1}, 0<p_{ij},\mu_i,\theta,\rho,\gamma<1, j=1,\ldots,m_i, i=1,\ldots,\ell.$$

We use both the Gibbs sampler and a random sampler to fit the model. The Gibbs sampler is used after collapsing over the p_{ij} and then samples are obtained from the posterior densities of the p_{ij} using the composition method (i.e., multiplication rule). Once samples of the p_{ij} are obtained, Bayesian predictive inference follows easily because $T_{ij}^{(1)}|p_{ij}\overset{ind}{\sim}\text{Binomial}(N_{ij}-n_{ij},p_{ij}), j=1,\ldots,m_i$, for the sampled clusters and $T_{ij}^{(2)}|p_{ij}\overset{ind}{\sim}\text{Binomial}(N_{ij},p_{ij}), j=1,\ldots,M_i$, for the nonsampled clusters. Given p_{ij}, $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ are independent. However, the Gibbs sampler is not easy to use because there are weakly identified parameters and this needs special attention. See TRN14 for the technical details and convergence monitoring.

For the random sampler, first note that conditionally a posteriori the p_{ij} are independent and

$$p_{ij}|s_{ij},\mu_i,\rho\overset{ind}{\sim}\text{Beta}\{s_{ij}+\mu_i(1-\rho)/\rho, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho)/\rho\}.$$

Accordingly, once samples are obtained from the joint posterior density of $\underset{\sim}{\mu},\theta,\rho,\gamma|\underset{\sim}{s}$, a sample of p_{ij} is easy to obtain. Then, after integrating out the p_{ij}, we have

$$\pi(\underset{\sim}{\mu},\theta,\rho,\gamma|\underset{\sim}{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\frac{B\{s_{ij}+\mu_i(1-\rho)/\rho, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho,(1-\mu_i)(1-\rho)/\rho\}}$$

$$\times\prod_{i=1}^{\ell}\frac{\mu_i^{\theta(1-\gamma)/\gamma-1}(1-\mu_i)^{(1-\theta)(1-\gamma)/\gamma-1}}{B\{\theta(1-\gamma)/\gamma,(1-\theta)(1-\gamma)/\gamma\}}\times\theta^{\alpha_o-1}(1-\theta)^{\beta_o-1}\rho^{\alpha_o-1}(1-\rho)^{\beta_o-1}$$

$$\gamma^{\alpha_o-1}(1-\gamma)^{\beta_o-1}, 0<\mu_i,\theta,\rho,\gamma<1, i=1,\ldots,\ell. \qquad (3)$$

See Appendix A for the more detailed computations using the random sampler. For the TIMSS data, the results from the Gibbs sampler and the random sample are similar (see TRN 14).

3. Numerical Analysis

We discuss an illustrative example using data from the Third International Mathematics and Science Study (TIMSS) and we perform a simulation study to confirm the superiority of the twofold small-area model. This section has three subsections.

In Subsection 3.1 we describe the model diagnostic procedures used for analysis. In Subsection 3.2 we analyze the TIMSS data. We compare the onefold and twofold models.

We have used the posterior mean (PM), posterior standard deviation (PSD), and 95% highest posterior density (HPD) interval to summarize the distributions. We also computed the numerical standard error (NSE), which is based on the batch means method; NSE is a measure of the repeatability of the entire sampling. In Subsection 3.3 we describe a simulation study.

3.1. *Model Diagnostics*

We discuss three goodness-of-fit procedures, the deviance information criterion (DIC) together with the complexity or effective number of parameters (PD), the conditional predictive ordinate (CPO) along with the logarithm of the pseudomarginal likelihood (LPML), and the Bayesian predictive *p*-value (BPP). The DIC, LPML, and BPP look at the overall fit of the model; see Gelman et al. (2013) for further discussions of these measures. We give expressions for the twofold model because it is easy to write down similar ones for the onefold model.

In the twofold model $s_{ij}|p_{ij}\overset{ind}{\sim}$Binomial$(n_{ij}, p_{ij})$, $p_{ij}|\mu_i\overset{ind}{\sim}$Beta$\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}$. Thus, integrating out the p_{ij} we get a product of beta-binomial probability mass functions,

$$p(\underset{\sim}{s}|\underset{\sim}{\mu},\rho)=\prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\binom{n_{ij}}{s_{ij}}\frac{B\{s_{ij}+\mu_i(1-\rho)/\rho, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho,(1-\mu_i)(1-\rho)/\rho\}}.$$

It is also true that $\mathrm{E}(\mathrm{s}_{ij}|\mu_i,\rho)=n_{ij}\mu_i$ and $\mathrm{Var}(s_{ij}|\mu_i,\rho)=n_{ij}\{1+(n_{ij}-1)\rho\}\mu_i(1-\mu_i)$.

Let

$$PD=\bar{D}-D(\underset{\sim}{\bar{\theta}},\ \bar{\gamma}),\quad DIC=\bar{D}+PD$$

respectively be the complexity of the model and the deviance information criterion, where $\bar{D}$ and $D(\underset{\sim}{\bar{\theta}},\ \bar{\gamma})$ are defined below for the onefold and twofold models.

Let $\mu_i^{(h)}, i=1,\ldots,\ell, \rho^{(h)}, h=1,\ldots,M$, denote the iterates of Gibbs sampling from the twofold model, $\bar{\mu}_i=\sum_{h=1}^{M}\mu_i^{(h)}/M, i=1,\ldots,\ell$, and $\bar{\rho}=\sum_{h=1}^{M}\rho^{(h)}/M$. Then, $D(\underset{\sim}{\bar{\mu}},\ \bar{\rho})=-2log\{\mathrm{p}(\underset{\sim}{\mathrm{s}}|\underset{\sim}{\bar{\mu}},\ \bar{\rho})\}$ and $\bar{D}=-2\sum_{h=1}^{M}log\{p(\underset{\sim}{s}|\underset{\sim}{\mu}^{(h)},\rho^{(h)})\}/M$.

Models with smaller DIC are preferred over models with larger DIC. Models are penalized both by the value of $\bar{D}$, which favors a good fit, and *PD*. Since $\bar{D}$ will decrease as the number of parameters in a model increases, *PD* compensates for this effect by favoring models with a smaller number of parameters. However, DIC tends to select overfitted models. The Bayesian predictive information criterion (BPIC) can protect against this effect but it is difficult to compute, it is not meant for dependent data, and consistency (as the sample size increases) is needed (see Ando 2007). The inconsistency problem can be overcome by integrating out the p_{ij} and the μ_i, but this creates dependent data.

Similar to the DIC, the second measure is the LPML. Both measures are based on the same cross-validation (leave-one-out) procedure. A summary statistic for CPO values is LPML; unlike the DIC, larger values of LPML indicate better fitting models (e.g., Geisser and Eddy 1979). For the twofold model the CPO is given by

$$\widehat{CPO}_{ij} = \left\{ \frac{1}{M} \sum_{h=1}^{M} \frac{1}{f\left(s_{ij}|p_{ij}^{(h)}\right)} \right\}^{-1}, j = 1, \ldots, m_i, i = 1, \ldots, \ell,$$

where $p_{ij}^{(h)}, h = 1, \ldots, M$, are the samples from $p_{ij}|s_{ij}, \mu_i, \rho$ and $s_{ij}|p_{ij}\overset{ind}{\sim}\text{Binomial}(n_{ij}, p_{ij})$. Again, it is interesting to note that for each (ij), $\widehat{CPO}_{ij}$ is the harmonic mean of the likelihoods $f(s_{ij}|p_{ij}^{(h)}), h = 1, \ldots, M$. Then,

$$LPML = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} log(\widehat{CPO}_{ij}).$$

The LPML, like the DIC, can discriminate between the onefold and the twofold models. We compute the CPO and the LPML at the cluster level, the LPML being preferable (easy to use).

Our third measure is the BPP for the two models. For the twofold model, the discrepancy function is

$$T_2(\underset{\sim}{s}; \underset{\sim}{\mu}, \rho) = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} \frac{\{s_{ij} - E(s_{ij}|\mu_i, \rho)\}^2}{Var(s_{ij}|\mu_i, \rho)}.$$

Then the BPP is $P\{T_2(\underset{\sim}{s}^{(rep)}; \underset{\sim}{\mu}, \rho) \geq T_2(\underset{\sim}{s}^{(obs)}; \underset{\sim}{\mu}, \rho)|\underset{\sim}{s}\}$, where probability is calculated over the iterates $(\underset{\sim}{\mu}^{(h)}, \rho^{(h)})$, $h = 1, \ldots, M$. Extremely small (near 0) or extremely large (near 1) values of this probability indicate that the model does not fit well.

3.2. *Illustrative Example*

First, we compare the two models using the three measures. For the onefold (twofold) model, $PD = 6.70$ ($PD = 7.98$), $DIC = 313$ ($DIC = 282$), $LPML = -609$ ($LPML = -575$), and $BPP = .000$ ($BPP = .467$). The BPP tells us that while the twofold model fits the TIMSS data reasonably well, the onefold model does not. The other two measures, *DIC* and *LPML*, tell us that the twofold model provides a better fit to the TIMSS data.

Using the onefold model, for θ $PM = .556$, $PSD = .052$, $NSE = .002$, and the 95% HPD interval is (.448,.654); for γ $PM = .112$, $PSD = .053$, $NSE = .001$, and the 95% HPD interval is (.034,.215). Using the twofold model, for θ $PM = .566$, $PSD = .055$, and $NSE = .002$, the 95% HPD interval is (.443,.662); for γ $PM = .078$, $PSD = .056$, $NSE = .002$, and the 95% HPD interval is (.001,.187). Thus inferences about θ and γ are very similar under the onefold and twofold models.

More importantly, the posterior mean of ρ is .217 with a standard deviation of .050, $NSE = .001$, and the 95% HPD interval of (.122,.309). This also shows that the twofold model, which accommodates the two-stage cluster sampling via the intracluster correlation, ρ, may be preferred.

In Table 2 we present posterior inference about the finite population proportions for the mathematics scores. We see that the posterior means of the onefold model can be larger or smaller than the posterior means of the twofold model. However, the posterior standard deviations for the twofold model are always larger than those of the onefold model. This clearly shows how the twofold model accommodates the clustering effect. In Table 2 we

Table 2. Comparison of posterior inference from the onefold and twofold models for the finite population proportions by areas for US students below average in mathematics

		Onefold			Twofold		
Area	Direct	PM	PSD	95% HPD	PM	PSD	95% HPD
NR	$.500_{.104}$	.515	.087	(.367, .696)	.528	.116	(.316, .747)
NO	$.316_{.067}$	.355	.063	(.234, .480)	.396	.093	(.234, .594)
NC	$.696_{.054}$	.682	.052	(.587, .785)	.667	.075	(.523, .806)
SR	$.649_{.068}$	.636	.061	(.527, .760)	.618	.087	(.440, .773)
SO	$.383_{.047}$	.395	.047	(.303, .483)	.410	.067	(.288, .539)
SC	$.827_{.047}$	.795	.048	(.694, .877)	.757	.071	(.617, .889)
CR	$.364_{.127}$	.427	.108	(.217, .609)	.459	.152	(.196, .717)
CO	$.542_{.093}$	.548	.080	(.386, .707)	.549	.111	(.336, .757)
CC	$.683_{.052}$	.667	.051	(.573, .766)	.660	.068	(.516, .778)
WR	$.296_{.078}$	.345	.076	(.203, .484)	.403	.107	(.203, .602)
WO	$.587_{.065}$	.582	.058	(.452, .683)	.583	.080	(.448, .751)
WC	$.703_{.037}$	.694	.036	(.618, .760)	.685	.051	(.591, .783)

NOTE: PM is the posterior mean, PSD is the posterior standard deviation and HPD is highest posterior density interval. The Monte Carlo errors of the posterior means are smaller than .004 in all cases, and in most cases are substantially smaller than .004. The direct estimate and its standard error are written as a_b where a is the direct estimate and b is its standard error.

have also presented the direct estimates. The direct estimates and their standard errors seem to be closer to the PMs and PSDs of the onefold model, but there are some differences (e.g., areas CR and WR).

In Figure 1 we present plots of the empirical posterior densities of the finite population proportions. These are obtained using the Parzen-Rosenblatt normal kernel density estimator with an optimal window width (e.g., Silverman 1986). In both pictures (onefold and twofold models) we observe a clear difference between the onefold and twofold models. The distributions under the twofold model are more spread out than those of the corresponding onefold model.

Using the TIMSS data (half sample) we perform two small empirical studies. First, we study the quality of the Bayesian predictive inference. Then the ‘true values’ of the finite population proportions (original sample) for the areas are .541, .347, .608, .600, .550, .667, .436, .421, .560, .458, .522, .643. Under the twofold model the 95% HPD interval of the finite population proportion of area SO misses the true value. But under the onefold model the 95% HPD intervals for areas SO, SC, CC miss the true value (see Table 1 for abbreviations). Thus, once again the twofold model provides a better fit than the onefold model.

Second, we investigate the effect of a larger number of areas. As our half-sample dataset has only twelve areas, we have artificially increased the number of areas. Specifically, we have bootstrapped the twelve areas in the half sample to fill in the additional number of areas to get 25, 50, 75, and 100 areas. Detailed comparisons between random sampling and Gibbs sampling are given in TRN14. For example, in the computations random sampling is twice as fast as Gibbs sampling, but the measures (e.g., DIC, LPML, and BPP) are similar.

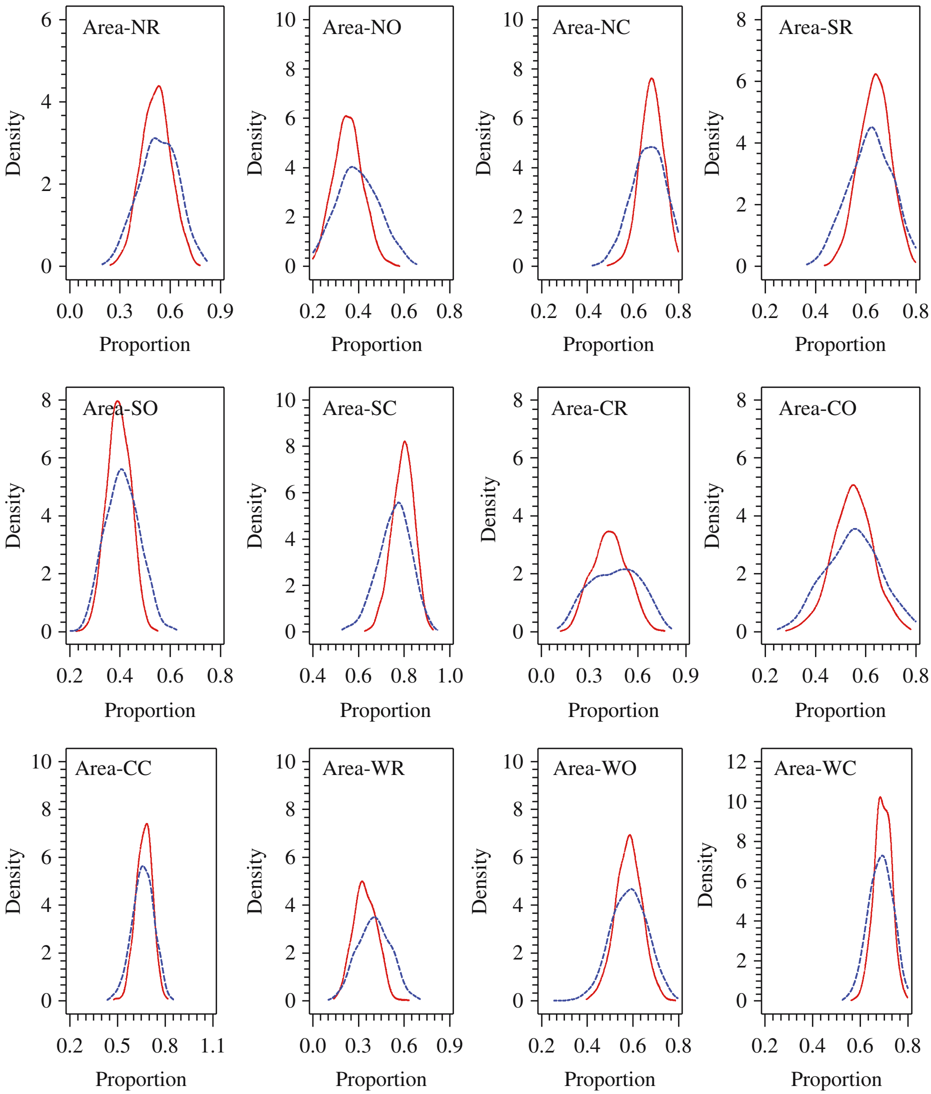


Fig. 1. Comparison of the onefold (solid) and twofold (dotted) models via posterior inferences of the finite population proportions of the empirical densities of finite population proportions by area

3.3. Simulation Study

We have performed a small simulation study to help understand how inferences about the finite population proportions change with the intracluster correlation coefficient (ρ) and the number (ℓ) of small areas. We have studied $\rho = .01, .10, .25, .50, .75$ and $\ell = 12, 25, 50, 75, 100$. Thus, there are twenty-five design points in our simulation study.

We have set the number of schools in each area to be 100 and the number of students within each school to be 15 (i.e., $N_{ij} = 15, j = 1, \ldots, M_i, \;\; M_i = 100, i = 1, \ldots, \ell$). We also hold $\theta = .60$ and $\gamma = .05$, near the posterior means calculated for the real data. We have taken a simple random sample of five schools from the 100 generated for the population, and a simple random sample of ten students from each selected school (i.e., $m_i = 5$ schools and $n_{ij} = 10$ students). So there are up to 100 areas each having 100 schools and each school having up to 15 students. So we have up to 10,000 schools and 150,000 students. The number of areas can be as large as current computing facilities allow because the area effects can be drawn using parallel computing via our method of random sampling (not Gibbs sampling).

We have simulated binary data from the twofold small-area model,

$$\mu_i | \theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{ \theta \frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma} \right\}, \; i = 1, \ldots, \ell,$$

$$p_{ij} | \mu_i, \rho \overset{ind}{\sim} \text{Beta}\left\{ \mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho} \right\}, \; j = 1, \ldots, M_i,$$

$$y_{ijk} | p_{ij} \overset{ind}{\sim} \text{Bernoulli}(p_{ij}), \; k = 1, \ldots, N_{ij}.$$

Thus, we have the true value of $P_i = \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} y_{ijk} / \sum_{j=1}^{M_i} N_{ij}$, $i = 1, \ldots, \ell$. We have taken 1,000 samples at each of the 25 design points.

In a similar way, we have generated data from the onefold model,

$$p_i | \theta, \gamma \overset{iid}{\sim} \text{Beta}\left\{ \theta \frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma} \right\}, \; j = 1, \ldots, M_i,$$

$$y_{ijk} | p_i \overset{ind}{\sim} \text{Bernoulli}(p_i), \; k = 1, \ldots, N_{ij},$$

with a subset of the same design points (i.e., $\rho = 0$).

For all generated data sets we fit the onefold and twofold models using random draws, as described for the computations. We have used parallel computing to fit the models. Note that we need to fit 25,000 simulated data sets.

Here, we have also studied the frequentist properties of our procedure. We compute the absolute bias (AB), relative absolute bias (RAB), and root posterior mean squared error (RPMSE). Specifically, we obtain $AB_{ih} = |PM_{ih} - P_{ih}|$, $RAB_{ih} = AB_{ih}/P_{ih}$ and $RPMSE_{ih} = \sqrt{PSD_{ih}^2 + AB_{ih}^2}$, $i = 1, \ldots, \ell$, $h = 1, \ldots, 1{,}000$. We have also computed the 95% HPD interval for each of the 1,000 simulated runs. We have looked at the width (W_{ih}) and the credible incidence (I_{ih}). Here $I_{ih} = 1$ if the 95% HPD interval contains the true value P_i and $I_{ih} = 0$ if the 95% credible interval does not contain the true value P_i. For each area and each design point we have taken the average of these quantities. For example, the estimated probability content of the 95% HPD interval for the i^{th} area is $C_i = \sum_{h=1}^{1000} I_{ih}/1{,}000$.

First, we discuss the simulations when data are generated from the twofold model. In Table 3 we present a comparison of the onefold and twofold models using these measures. The coverages for the twofold model are much closer to the nominal value of 95% than those from the onefold model. In some cases the coverages from the onefold model are much too small. However, the 95% HPD intervals from the twofold model are wider than those from the onefold model. These effects are much larger as ρ increases for each ℓ, thereby clearly showing how the twofold model takes care of the clustering effect. All measures (AB, RAB, RPMSE) for the twofold model are smaller than those for the

Table 3. Simulation for data drawn from the twofold model: Comparison of coverage and widths of 95% HPD intervals and absolute bias, relative absolute bias and, root posterior mean squared error at twelve design points

ℓ	ρ	Model	C-HPD	W-HPD	AB	RAB	RPMSE
25	.10	TFM	$.940_{.0015}$	$.276_{.0002}$	$.056_{.0003}$	$.098_{.0005}$	$.096_{.0002}$
		OFM	$.860_{.0022}$	$.227_{.0001}$	$.061_{.0003}$	$.106_{.0005}$	$.090_{.0002}$
	.25	TFM	$.938_{.0016}$	$.318_{.0002}$	$.068_{.0003}$	$.121_{.0007}$	$.113_{.0002}$
		OFM	$.732_{.0029}$	$.227_{.0001}$	$.081_{.0004}$	$.142_{.0008}$	$.107_{.0003}$
	.50	TFM	$.918_{.0018}$	$.355_{.0003}$	$.077_{.0004}$	$.136_{.0008}$	$.127_{.0003}$
		OFM	$.612_{.0031}$	$.227_{.0002}$	$.107_{.0005}$	$.184_{.0009}$	$.129_{.0004}$
	.75	TFM	$.944_{.0015}$	$.417_{.0003}$	$.083_{.0004}$	$.147_{.0009}$	$.145_{.0003}$
		OFM	$.495_{.0032}$	$.222_{.0003}$	$.137_{.0006}$	$.239_{.0012}$	$.157_{.0006}$
50	.10	TFM	$.940_{.0011}$	$.273_{.0001}$	$.058_{.0002}$	$.104_{.0004}$	$.097_{.0001}$
		OFM	$.857_{.0016}$	$.225_{.0001}$	$.060_{.0002}$	$.105_{.0004}$	$.090_{.0002}$
	.25	TFM	$.935_{.0011}$	$.314_{.0002}$	$.067_{.0002}$	$.119_{.0005}$	$.112_{.0002}$
		OFM	$.727_{.0020}$	$.228_{.0001}$	$.082_{.0003}$	$.143_{.0005}$	$.108_{.0002}$
	.50	TFM	$.936_{.0011}$	$.350_{.0002}$	$.074_{.0002}$	$.133_{.0005}$	$.124_{.0002}$
		OFM	$.607_{.0022}$	$.229_{.0001}$	$.108_{.0004}$	$.190_{.0007}$	$.131_{.0003}$
	.75	TFM	$.942_{.0010}$	$.386_{.0002}$	$.080_{.0003}$	$.143_{.0006}$	$.135_{.0002}$
		OFM	$.492_{.0022}$	$.222_{.0002}$	$.137_{.0004}$	$.240_{.0008}$	$.157_{.0004}$
100	.10	TFM	$.946_{.0007}$	$.275_{.0001}$	$.056_{.0001}$	$.100_{.0003}$	$.096_{.0001}$
		OFM	$.862_{.0011}$	$.225_{.0001}$	$.060_{.0001}$	$.105_{.0003}$	$.089_{.0001}$
	.25	TFM	$.939_{.0008}$	$.311_{.0001}$	$.066_{.0002}$	$.117_{.0003}$	$.110_{.0001}$
		OFM	$.752_{.0014}$	$.229_{.0001}$	$.080_{.0002}$	$.140_{.0004}$	$.106_{.0002}$
	.50	TFM	$.930_{.0008}$	$.340_{.0001}$	$.075_{.0002}$	$.136_{.0004}$	$.122_{.0001}$
		OFM	$.602_{.0015}$	$.227_{.0001}$	$.109_{.0003}$	$.192_{.0005}$	$.131_{.0002}$
	.75	TFM	$.936_{.0008}$	$.385_{.0001}$	$.082_{.0002}$	$.150_{.0005}$	$.137_{.0001}$
		OFM	$.485_{.0016}$	$.222_{.0001}$	$.140_{.0003}$	$.248_{.0006}$	$.160_{.0003}$

NOTE: TFM is the twofold model and OFM is the onefold model. W-HPD and C-HPD are respectively the width and the probability content of a HPD interval. A, AB, and RPMSE are the absolute bias, relative absolute bias and root posterior mean square error. The notation a_b means that a is the estimate and b is the standard error.

onefold model. These effects become more intense for larger ρ. Again this shows the superiority of the twofold model over the onefold model.

In Table 4 we present summaries of PD, DIC, LPML, and BPP. As expected, the PDs for the twofold model should be larger than those for the onefold model. All the DICs for the twofold model are smaller than the corresponding ones for the onefold model, and this disparity becomes larger as ℓ and ρ increase. The results are the same for the LPML. Under the onefold model most of the BPPs are near 0, but under the twofold model the corresponding BPPs are around 0.5. These measures show that while the twofold model is more complex, it is superior to the onefold model. In TRN14 we compare plots of the sample distributions of the negative LPML under the onefold and twofold models over the 1,000 runs by ℓ and ρ. The negative LPML under the twofold model are smaller than under the onefold model and this discrepancy increases with both ρ and ℓ. There are overlaps of distributions when $\rho = .10$ but not for other values of ρ.

Second, we discuss the simulations when data are generated from the onefold model. In Table 5 we present comparisons of the onefold and twofold models. As expected, the onefold model is slightly better than the twofold model. AB, RAB, RPMSE are only

Table 4. Summaries of the 1,000 simulation runs with (data drawn from the twofold model) for the complexity, deviance information criterion, log pseudomarginal likelihood, and the Bayesian predictive p-value by ℓ, ρ and model

		Onefold				Twofold			
ℓ	ρ	PD	DIC	LPML	BPP	PD	DIC	LPML	BPP
25	.10	4.998	553	−1107	.010	15.61	521	−1086	.462
	.25	6.795	625	−1190	.000	12.09	557	−1087	.467
	.50	9.289	717	−1333	.000	9.38	530	−1038	.478
	.75	10.970	710	−1465	.000	9.27	419	−953	.479
50	.10	4.980	1141	−2303	.000	30.86	1081	−2261	.461
	.25	6.981	1291	−2475	.000	23.60	1156	−2260	.473
	.50	9.566	1450	−2773	.000	17.29	1107	−2162	.479
	.75	11.220	1465	−3050	.000	15.06	870	−1983	.495
100	.10	5.015	2278	−4606	.000	61.67	2162	−4522	.472
	.25	6.889	2574	−4943	.000	45.68	2316	−4524	.476
	.50	9.649	2873	−5526	.000	31.90	2200	−4320	.485
	.75	11.420	2873	−6104	.000	28.39	1720	−3957	.496

NOTE: PD is the effective number of parameters, DIC is the deviance information criterion, LPML is the log pseudomarginal likelihood and BPP is the Bayesian predictive *p*-value based on the chi-squared measure. The standard errors are negligible.

slightly smaller under the onefold model. However, the coverages of the HPD intervals under the twofold model are closer to the nominal value of 95%, with those from the onefold model being slightly smaller. This is due to the phenomenon that the intervals under the onefold model are narrower.

In Table 6 we present summaries of PD, DIC, LPML, and the BPP. These measures are very similar for the two models. While the BPPs are different, they show that the two models fit equally well. However, the main difference is in PD, the complexity of the model. While the twofold model is more complex than the onefold model, they fit equally well when the onefold model is expected to hold.

Table 5. Simulation for data drawn from the onefold model: Comparison of coverage and widths of 95% HPD intervals and absolute bias, relative absolute bias, and root posterior mean squared error

ℓ	Model	C-HPD	W-HPD	AB	RAB	RPMSE
12	TFM	$.953_{.0019}$	$.244_{.0002}$	$.048_{.0003}$	$.085_{.0007}$	$.084_{.0002}$
	OFM	$.933_{.0023}$	$.223_{.0002}$	$.047_{.0003}$	$.084_{.0006}$	$.079_{.0002}$
25	TFM	$.949_{.0014}$	$.234_{.0001}$	$.049_{.0002}$	$.088_{.0005}$	$.082_{.0002}$
	OFM	$.929_{.0017}$	$.219_{.0001}$	$.046_{.0002}$	$.083_{.0005}$	$.078_{.0002}$
50	TFM	$.948_{.0010}$	$.233_{.0001}$	$.048_{.0002}$	$.086_{.0003}$	$.082_{.0001}$
	OFM	$.943_{.0010}$	$.220_{.0001}$	$.047_{.0002}$	$.083_{.0003}$	$.078_{.0001}$
75	TFM	$.954_{.0008}$	$.236_{.0001}$	$.046_{.0001}$	$.082_{.0003}$	$.081_{.0001}$
	OFM	$.945_{.0008}$	$.221_{.0000}$	$.045_{.0001}$	$.080_{.0002}$	$.077_{.0001}$
100	TFM	$.959_{.0006}$	$.236_{.0001}$	$.046_{.0001}$	$.081_{.0002}$	$.081_{.0001}$
	OFM	$.948_{.0007}$	$.221_{.0000}$	$.045_{.0001}$	$.079_{.0002}$	$.077_{.0001}$

NOTE: TFM is the twofold model and OFM is the onefold model. W-HPD and C-HPD are respectively the width and probability content of a HPD interval. A, AB and RPMSE are the absolute bias, relative absolute bias, and root posterior mean square error. The notation a_b means that a is the estimate and b is the standard error.

Table 6. Summaries of the 1,000 simulation runs (data are drawn from the onefold model) for the complexity, deviance information criterion, log pseudomarginal likelihood, and the Bayesian predictive p-value by ℓ and model

	Onefold				Twofold			
ℓ	PD	DIC	LPML	BPP	PD	DIC	LPML	BPP
12	3.673	250	−530	.445	8.94	233	−531	.661
25	3.362	486	−1055	.553	16.57	460	−1059	.781
50	3.587	1023	−2203	.479	34.43	962	−2208	.772
75	3.574	1527	−3301	.526	52.81	1438	−3308	.852
100	3.692	2043	−4401	.538	70.73	1914	−4410	.882

NOTE: PD is the effective number of parameters, DIC is the deviance information criterion, LPML is the log pseudomarginal likelihood, and BPP is the Bayesian predictive *p*-value based on the chi-squared measure. The standard errors are negligible.

4. Concluding Remarks

We have developed a twofold hierarchical Bayesian model to analyze binary data arising from a twofold sample design for small areas. This model incorporates an intracluster correlation, and it is an extension of the two-stage hierarchical Bayesian model of Nandram and Sedransk (1993) and, more importantly, the twofold model of Stukel and Rao (1997; 1999) for binary data. A onefold model ignores the intracluster correlation. We have performed a Bayesian predictive inference for the finite population proportion of each area. We have discussed how to study the onefold and twofold small-area models in detail. As an illustrated example, we have used synthetic data from TIMSS, a study of the performance of US students at the third grade in mathematics. We have also performed a simulation study to compare the onefold and twofold models. We have shown how to overcome a difficulty in running the Gibbs sampler that we initially used to fit the twofold model (see TRN14).

We have shown that when there is clustering within each area, the onefold model gives poor performance, and the twofold model is much more preferable. The onefold model can lead to estimators that differ from the twofold model in terms of both location and spread. Our simulation study provides strong evidence that the twofold model is to be preferred when there is a two-stage cluster sampling design within each area. This is a direct consequence of the effect of the intracluster correlation. The Bayesian measures (deviance information criterion, log pseudomarginal likelihood, Bayesian predictive *p*-value) and frequentist measures (bias, mean squared error, coverage) show that the twofold model is better than the onefold model. While we have demonstrated that the twofold model is preferred when data are available from a twofold sampling design with cluster sampling, other sampling designs (e.g., stratification) in each area will give different results, and these need to be investigated separately.

We have shown that the twofold model is preferable to the onefold model for the TIMSS data. Although the two models give similar results, we have better point and interval estimates from the twofold model. We can see from Table 2 that there are some possibly interesting findings for TIMSS data even though we have not used all features of the data.

Apparently a school in a western rural (WR) area is the best and city schools (NC, SC, CC, WC) are not so good.

This research has opened up many avenues for future work on twofold small-area models. First, for a more realistic analysis of the TIMSS data, it is possible to incorporate the survey weights into our analysis. Second, it may be desirable to have the intracluster correlation to vary with area. It is expected that the computation will be challenging because with a single intracluster correlation there is long-range dependence among the iterates from the Gibbs sampler. Third, it is desirable to study threefold models (states within regions and counties within states). Fourth, we can look at polychotomous data instead of binary data; in TIMSS one can use three levels for mathematics score (below average, average, above average). Fifth, we can consider multivariate binary data; in TIMSS there are both mathematics and science scores. This will lead naturally to consider test of independence for two categorical variables. Sixth, benchmarking for small areas is also an important problem (states within regions and counties within states). Seventh, we can look at covariates via logistic regression; in TIMMS there are covariates. Eight, we can use nonparametric models (e.g., Dirichlet process mixtures and mixture of finite Polya trees) to help robustify our twofold model.

APPENDIX A: Computation Without Gibbs Sampling

Long-range dependence is a general problem for the hierarchical Bayesian model when Markov chain Monte Carlo methods are used to fit it. Typically long-range dependence is due to weak identifiability in some parameters and/or indirect functional relation among the parameters, and this causes poor mixing in the Gibbs sampler. The solution of thinning the iterates, used in practice, is not really efficient. These problems occur when the twofold model is fitted, and so it is pertinent to present an alternative algorithm that uses just random samples.

Our strategy is to use the composition method (i.e., multiplication rule) to draw random samples from the posterior density $\pi(\underset{\sim}{\mu}, \theta, \rho, \gamma|\underset{\sim}{y})$. That is,

$$\pi(\underset{\sim}{\mu}, \theta, \rho, \gamma|\underset{\sim}{y}) = \left\{\prod_{i=1}^{\ell} \pi(\mu_i|\theta, \rho, \gamma, \underset{\sim}{y})\right\} g(\theta, \rho, \gamma|\underset{\sim}{y}).$$

Integrating out $\mu_i, i = 1, \ldots, \ell$, the joint posterior density of $\theta, \rho, \gamma|\underset{\sim}{y}$ is

$$\pi(\theta, \rho, \gamma|\underset{\sim}{y}) = A\left[\prod_{i=1}^{\ell}\left\{\int_0^1 g_i(\mu_i)f(\mu_i)d\mu_i\right\}\right]\theta^{\alpha_o-1}(1-\theta)^{\beta_o-1}\rho^{\alpha_o-1}(1-\rho)^{\beta_o-1}\gamma^{\alpha_o-1}$$

$$(1-\gamma)^{\beta_o-1},$$

where A is a normalization constant hence forth omitted,

$$g_i(\mu_i) = \prod_{j=1}^{m_i} \frac{B\{s_{ij} + \mu_i(1-\rho)/\rho, n_{ij} - s_{ij} + (1-\mu_i)(1-\rho)/\rho\}}{B\{\mu_i(1-\rho)/\rho, (1-\mu_i)(1-\rho)/\rho\}},$$

and

$$f(\mu_i) = \frac{\mu_i^{\theta(1-\gamma)/\gamma-1}(1-\mu_i)^{(1-\theta)(1-\gamma)/\gamma-1}}{B\{\theta(1-\gamma)/\gamma, (1-\theta)(1-\gamma)/\gamma\}}.$$

Note that while $g_i(\mu_i)$ is the ratio of two beta functions (computations discussed earlier) both of which are functions of ρ but not θ and $\gamma, f(\mu_i)$ is a function of θ and γ but not ρ. More importantly, $f(\mu_i)$ is a density function of a beta random variable. We can integrate out the μ_i, one at a time, and form their product to obtain the complete integral. Thus, we only need to discuss how to compute $\int_0^1 g_i(\mu_i)f(\mu_i)d\mu_i$, $i = 1, \ldots, \ell$, for one area. Also, note that $f(\mu_i)$ does not depend on i under the integral sign. While this integral can be computed using Monte Carlo methods, it is much more efficient to use numerical integration in the following way.

Let $F(\cdot)$ denote the cdf corresponding to $f(\cdot)$. Partition the interval (0,1) into a mesh of G subintervals $[a_0, a_1], [a_1, a_2], \ldots, [a_{G-1}, a_G]$ where $a_0 = 0$, $a_i = i/G, i = 1, \ldots, G$. Then, using the Riemann middle sum, it is easy to show that

$$\lim_{G\to\infty} \sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F(a_v) - F(a_{v-1})\} = \int_0^1 g_i(x)f(x)dx, \; i = 1, \ldots, \ell.$$

Thus, for reasonably large G, $\sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F(a_v) - F(a_{v-1})\} \approx \int_0^1 g_i(x)f(x)\,dx$, $i = 1, \ldots, \ell$.

Together with integrating out the μ_i, we have also integrated out θ, ρ, where we use Gaussian quadrature via Legendre orthogonal polynomials,

$$p(\gamma|\underset{\sim}{y}) \approx \sum_{g_1=1}^{G}\sum_{g_2=1}^{G} w_{g_1}w_{g_2}\left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, x_{g_1}, x_{g_2}, \gamma|\underset{\sim}{y})d\mu_i\right\},$$

where $w_g, g = 1, \ldots, G$, are the weights and $x_g, g = 1, \ldots, G$, are roots of the Legendre polynomial with x_{g_1} and x_{g_2} corresponding to θ and ρ respectively. Note that the single integral over each μ_i is done as described above and the whole procedure is a three-dimensional integral. Now, using univariate grids, samples of the posterior density of γ are obtained in exactly the same manner as described for the onefold model using the univariate grid sampler.

Then, conditional on γ, the posterior density of ρ is

$$p(\rho|\gamma, \underset{\sim}{y}) \approx \sum_{g=1}^{G} w_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, x_g, \rho|\gamma, \underset{\sim}{y})d\mu_i\right\}.$$

Again using the univariate grid sampler, samples are drawn from the posterior density of ρ.

Next, conditional on (ρ, γ), the posterior density of θ is

$$p(\theta|\rho, \gamma, \underset{\sim}{y}) \approx \left\{\prod_{i=1}^{\ell}\int_0^1 \pi(\mu_i, \theta|\rho, \gamma, \underset{\sim}{y})d\mu_i\right\}.$$

Again using the univariate grid sampler, samples are drawn from the posterior density of θ.

Finally, conditional on (θ, ρ, γ), the μ_i are independent and samples are again obtained from $\pi(\mu_i | \theta, \rho, \gamma, \underset{\sim}{y})$ using the univariate grid sampler. We have always used 100 grids for the μ_i, θ, ρ and γ.

APPENDIX B: A Multistage Hierarchical Bayesian Model

In TIMSS the countries can be compared, a task beyond the scope of the current article. The small areas (regions and communities) are clustered within the countries and the schools are clustered within these small areas. This is a generalization of the twofold design, which we have discussed in detail in Section 2, to a threefold design. Thus we describe the multistage model mainly for reasons of theoretical interest.

The multifold hierarchical Bayesian model is

$$y_{ij_1,\ldots,j_k} | \mu_{ij_1,\ldots,j_{k-1}} \overset{ind}{\sim} \text{Bernoulli}(\mu_{ij_1,\ldots,j_{k-1}}).$$

For $s = 1,\ldots,k-1$,

$$\mu_{ij_1,\ldots,j_{k-s}} | \mu_{ij_1,\ldots,j_{k-(s+1)}}, \gamma_1 \overset{ind}{\sim} \text{Beta}\left\{\mu_{ij_1,\ldots,j_{k-(s+1)}} \frac{1-\gamma_1}{\gamma_1},\ (1-\mu_{ij_1,\ldots,j_{k-(s+1)}}) \frac{1-\gamma_1}{\gamma_1}\right\}.$$

and

$$\mu_i | \theta, \gamma_k \overset{iid}{\sim} \text{Beta}\left\{\theta \frac{1-\gamma_k}{\gamma_k},\ (1-\theta)\frac{1-\gamma_k}{\gamma_k}\right\}.$$

Finally, a priori

$$\theta, \gamma_1, \ldots, \gamma_k \overset{iid}{\sim} \text{Uniform}(0,1).$$

Note that in this hierarchical Bayesian model, the first two stages are conjugate and the other stages are nonconjugate. More importantly, the correlation between two units at the first stage is γ_1. Furthermore, when the first-stage means are integrated out, the correlation between two units in two different clusters is γ_2, and so on. It is expected that the correlations will decay as we go down the hierarchical structure of the model. That is, the correlation between two units at the area level is expected to be the smallest while the correlation at the last stage of the multistage cluster sampling design is expected to be the largest.

While the multistage model is of practical importance, it would need significant research to develop it into a useful methodology and it is expected that the computation will be challenging.

5. References

Ando, T. 2007. “Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models.” *Biometrika* 94: 443–458. Doi: http://dx.doi.org/10.1093/biomet/asm017.

Bedrick, E.J. 1983. “Adjusted Chi-Squared Tests for Cross-Classified Tables of Survey Data.” *Biometrika* 70: 591–595. Doi: http://dx.doi.org/10.1093/biomet/70.3.591.

Brier, S.S. 1980. “Analysis of Contingency Tables Under Cluster Sampling.” *Biometrika* 67: 591–596. Doi: http://dx.doi.org/10.1093/biomet/67.3.591.

Calsyn, C., P. Gonzales, and M. Frase. 1999. “Highlights from TIMSS.” National Center for Education Statistics, Washington, DC. Doi: http://mces.ed.gov/timss.

Datta, G.S. and M. Ghosh. 1991. “Bayesian Prediction in Linear Models: Applications to Small Area Estimation.” *Annals of Statistics* 19: 1748–1770.

Foy, P., K. Rust, and A. Schleicher. 1996. “Sample Design.” In *TIMMS Technical Report, Volume I: Design and Development*, edited by M.O. Martin and D.L. Kelly, pagenumber. Chestnut Hill, MA: Boston College.

Fuller, W.A. and G.E. Battese. 1973. “Transformations for Estimation of Linear Models with Nested-Error Structure.” *Journal of the American Statistical Association* 68: 626–632. Doi: http://dx.doi.org/10.1080/01621459.1973.10481396.

Gelfand, A., D. Dey, and H. Chang. 1992. “Model Determination using Predictive Distributions with Implementation via Sampling-based Methods.” In *Bayesian Statistics* 4, 147–168. New York: Oxford University Press.

Geisser, S. and W. Eddy. 1979. “A Predictive Approach to Model Selection.” *Journal of the American Statistical Association* 74: 153–160. Doi: http://dx.doi.org/10.1080/01621459.1979.10481632.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. *Bayesian Data Analysis*, 3rd ed. New York: Chapman & Hall/CRC.

Ghosh, M. and P. Lahiri. 1988. “Bayes and Empirical Bayes Analysis in Multistage Sampling.” In *Statistical Decision Theory and Related Topics IV*, Vol. 1, edited by S.S. Gupta and J.O. Berger. 195–212. New York: Springer.

Hamilton, J. 2009. *President Obama, U.S. Secretary of Education Duncan Announce National Competition to Advance School Reform*. U.S. Department of Education: Available at: http://www.ed.gov/news/pressreleases/2009/07/07242009.html.

Holt, D., A.J. Scott, and P.D. Ewings. 1980. “Chi-Squared Tests with Survey Data.” *Journal of the Royal Statistical Society, Series A* 143: 303–320. Doi: http://dx.doi.org/10.2307/2982131.

Malec, D. and J. Sedransk. 1985. “Bayesian Inference for Finite Population Parameters in Multistage Cluster Sampling.” *Journal of the American Statistical Association* 80: 897–902. Doi: http://dx.doi.org/10.1080/01621459.1985.10478200.

Molina, I., B. Nandram, and J.N.K. Rao. 2014. “Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach.” *Annals of Applied Statistics* 8: 852–885. Doi: http://dx.doi.org/10.1214/13-AOAS702.

Nandram, B. 2014. *Bayesian Predictive Inference for a Proportion Under a Two-Fold Small Area Model*. Technical Report, Department of Mathematical Sciences, Worcester Polytechnic Institute, 1–43. (Available on request.)

Nandram, B., D.R. Bhatta, J. Sedransk, and D. Bhadra. 2013. “A Bayesian Test of Independence in a Two-Way Contingency Table Using Surrogate Sampling.” *Journal of Statistical Planning and Inference* 143: 1392–1408. Doi: http://dx.doi.org/10.1016/j.jspi.2013.03.011.

Nandram, B. 1998. “A Bayesian Analysis of the Three-Stage Hierarchical Multinomial Model.” *Journal of Statistical Computation and Simulation* 61: 97–126. Doi: http://dx.doi.org/10.1080/00949659808811904.

Nandram, B. and J. Sedransk. 1993. "Bayesian Predictive Inference for a Finite Population Proportion: Two-Stage Cluster Sampling." *Journal of the Royal Statistical Society, Series B* 55: 399–408.

Natarajan, R. and R.E. Kass. 2000. "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association* 95: 227–237. Doi: http://dx.doi.org/10.1080/01621459.2000.10473916.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: Wiley.

Rao, J.N.K. and A.J. Scott. 1981. "The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-Way Tables." *Journal of the American Statistical Association* 76: 221–230. Doi: http://dx.doi.org/10.1080/01621459.1981.10477633.

Rao, J.N.K. and A.J. Scott. 1984. "On Chi-Squared Tests for Multi-way Tables with Cell Proportions Estimated from Survey Data." *Annals of Statistics* 12: 46–60.

Scott, A.J. and D. Holt. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of the American Statistical Association* 77: 848–854. Doi: http://dx.doi.org/10.1080/01621459.1982.10477897.

Scott, A. and T.M.F. Smith. 1969. "Estimation in Multi-Stage Surveys." *Journal of the American Statistical Association* 101: 1387–1397. Doi: http://dx.doi.org/10.1080/01621459.1969.10501015.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Stukel, D.M. and J.N.K. Rao. 1997. "Estimation of Regression Models with Nested Error Regression Structure and Unequal Error Variances Under Two and Three Stage Cluster Sampling." *Statistics & Probability Letters* 35: 401–407. Doi: http://dx.doi.org/10.1016/S0167-7152(97)86602-3.

Stukel, D.M. and J.N.K. Rao. 1999. "On Small-Area Estimation Under Two-Fold Nested Error Regression Models." *Journal of Statistical Planning and Inference* 78: 131–147. Doi: http://dx.doi.org/10.1016/S0378-3758(98)00211-0.

Toto, M.C.S. and B. Nandram. 2010. "A Bayesian Predictive Inference for Small Area Means Incorporating Covariates and Sampling Weights." *Journal of Statistical Planning and Inference* 140: 2963–2979. Doi: http://dx.doi.org/10.1016/j.jspi.2010.03.043.

Yan, G. and J. Sedransk. 2007. "Bayesian Diagnostic Techniques for Detecting Hierarchical Structure." *Bayesian Analysis* 2: 735–760. Doi: http://dx.doi.org/10.1214/07-BA230.

Yan, G. and J. Sedransk. 2010. "A Note on Bayesian Residuals as a Hierarchical Model Diagnostic Technique." *Statistical Papers* 51: 1–10. Doi: http://dx.doi.org/10.1007/s00362-007-0111-2.

Received December 2012
Revised September 2015
Accepted October 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 209–229, http://dx.doi.org/10.1515/JOS-2016-0010

SELEKT – A Generic Tool for Selective Editing

Anders Norberg[1]

The aim of selective editing is to make the often resource-demanding traditional editing process in business surveys more effective without a substantial loss in the precision of the output statistics. Recently, Statistics Sweden has developed a generic software package for selective editing called SELEKT. The method underpinning the software promotes continuous measurement of the suspicion of error response rather than a dichotomous measure using traditional edits. SELEKT is flexible and can be used in the production of surveys with different designs. Business surveys have diverse output regarding the number of variables, statistical measures, and domains of study definitions. The key objective of selective editing is to rank suspected errors in data according to the anticipated impact on the output. The software therefore has options to set different weights for different parts of the output to meet the needs of the main users of the statistics. Statistics Sweden has implemented SELEKT in eleven surveys to date. The experience gained will be used to provide recommendations on how to perform selective editing. This article will give an insight into SELEKT and its underlying theoretical base.

Key words: Edit rule; anticipated value; suspicion; impact; score function.

1. Introduction

The aim of statistical data editing is to detect and adjust errors in data resulting from collection and processing. It is considered a necessary survey operation because errors in survey data may distort estimates, complicate further processing, and decrease provider and user confidence (Granquist 1984). The emphasis of the editing task is slowly moving from just cleaning up the data, though this remains a necessary operation, to identifying and collecting data on errors, problem areas, and the causes of error to provide a basis for a continuous improvement of the whole survey vehicle (Granquist 1997). A further goal is to provide information for the quality declaration of the output statistics.

Editing appears in various forms in the business survey process, from respondents entering information into electronic questionnaires to the final checking of results prior to publication. Methods and procedures are mainly divided into micro- and macroediting. Microediting means the checking of individual data records, preferably as soon as data is

[1] Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden. Email: anders.norberg@scb.se. For questions regarding SELEKT, contact Karin Lindgren, Email: karin.lindgren@scb.se.

Acknowledgments: Gunnar Arvidson managed the project that specified the method for selective data editing useful for National Statistical Institutes. Chandra Adolfsson, Peter Gidlund and Lennart Nordberg were permanent members of the project team. Anders Kraftling and Can Tongur produced a SAS® application. Karin Lindgren made several implementations at Statistics Sweden. Everybody mentioned was instrumental in the development and implementation of SELEKT. The author is grateful to the assistant editor and referees for their constructive comments.

available. Checking rules are logical conditions or restrictions applied in order to check the validity, internal consistency, and plausibility of individual units' data. A routine flags data that fail the checking rule and enables the analyst, who performs the manual editing, to change data interactively. The analyst must have knowledge of the survey, the population and the kind of errors that are likely to occur. The flagged data can be compared to reference data such as data on the same unit from previous survey rounds, data on similar units, and data from an external register or information on the internet. Finally, the editor may recontact a respondent to check whether a value is declared as correct or to obtain a new value for a variable that was originally incorrect or was suspected to be incorrect. Procedures like this are called manual editing, interactive editing, production editing or simply microediting. Macroediting is performed when all, or most of, the data has been collected. The so-called distribution method checks for outliers in the final data set, as microediting does, but now the comparisons with similar units are more in focus. The aggregation method analyses statistical output, in order to check that no influential errors in microdata remain and that no processing error has been introduced.

Particularly in business surveys, editing is recognised as a time- and resource-consuming survey process, especially when recontacts are necessary. The costs include not only financial and human resources, but also loss in timeliness and excessive respondent burden. Further, there is a danger of distorting true values to fit them to preconceived models. This "overediting" gives users a false sense of security as far as data quality is concerned (Granquist and Kovar 1997).

New theories and methods to reduce the resources spent on editing survey data have been developed over the last thirty years or so. The leading idea is to concentrate resources in microediting on observations that affect the estimates, accepting that final data sets do contain errors with a negligible effect on the statistics produced (Granquist and Kovar 1997). Several methods have been implemented for this purpose. An early method, proven successful by experience, was presented by Hidiroglou and Berthelot (1986). This method considers both the level and the relative change from a previous survey round for a survey variable. Robustness is achieved by using the median and quartiles in the analysis of data.

In the early 1990s, methods based on a score function for selective data editing, henceforth abbreviated SE, emerged. In these methods, survey units that fail at least one edit rule are ranked by the score in order to give priority to those units that have the largest overall anticipated effect on the statistics produced. SE yields a global score for a primary sampling unit, for all the data delivered for that unit, that is, for any cluster elements and for one or many variables. The purpose of SE is merely to reduce the cost of the manual follow-up work without a substantial loss in the precision of the output statistics. Reducing manual follow-up with recontacts also lightens the workload of respondents. Macroediting by the aggregate method and SE have apparent similarities in that both methods focus on the set of estimates to be published. It turns out that early manual editing by SE can reduce the late macroediting.

De Waal et al. (2011) emphasise the detection and correction of systematic errors as a first step in an editing process. A systematic error occurs frequently between responding units when they misunderstand or misread a survey question in the same way. Causes of major systematic errors can be discovered through analysis of edit failures in data; frequent failures of an edit rule are indications of a problem for the respondents.

Data values with small systematic errors are difficult to distinguish from true values as they lie in the interior of a statistical distribution. Causes of these so-called inliers must be found by other methods than editing and the problems must be solved by improved data collection instruments. SE methods, leaving a part of data without manual follow-up, are not appropriate when the survey suffers from severe systematic errors.

The Australian Bureau of Statistics, ABS, has developed a form of selective editing called “significance editing”, discussed in Latouche and Berthelot (1992), Lawrence and McDavitt (1994), Farwell and Raine (2000), Lawrence and McKenzie (2000), and Farwell (2004). A basic significance editing score is a prediction of the change in an estimate due to correcting reporting errors, as it is an estimate of the reduction in reporting bias. If such a score is not possible to approximate, a score which is correlated to the expected reduction in reporting bias should be used. The approach has resulted in a noticeable improvement in editing efficiency. ABS has developed the tool Significance Editing Engine, SigEE, presented by Brinkley et al. (2011).

Di Zio and Guarnera (2013) present the generic tool Selemix, developed by the Italian National Institute of Statistics, Istat. They assume a normal model for the true data, also possible in log scale, and an “intermittent” error mechanism such that a proportion of data is contaminated by an additive Gaussian error. Based on these assumptions, a latent class model is used to derive the distribution of “true” data conditional on observed data. This approach allows scores to be interpreted as anticipated impact of errors and allows a selective editing procedure to be defined that identifies units containing the errors that have the largest influence on the estimates of interest.

A series of projects was started at Statistics Sweden, SCB, in 2006 with the main purpose of analysing which methods should be recommended for editing and for construction of a generic IT tool. Case studies focused on how to use SE (Adolfsson and Gidlund 2008). Nine of the most editing-intensive surveys were included in the project. The case studies show that SE will lead to efficiency gains and likely cost reductions in many surveys. The implementation of SE demands intensive testing in every specific survey. The variation between the surveys regarding survey design, data structure, output statistics, and so on, is large. A generic tool for editing must therefore be very flexible to deal with these differences. Efficiency can also be increased by dealing with known or encountered measurement problems in the auditing and picking low-hanging fruit that would improve the surveys.

The next editing project documented the best methods used in the case studies for various situations. Key algorithms for scores and the aggregation thereof were the framework.

The third phase was developing an IT tool. SELEKT is SAS® application for SE which establishes a general solution that can be implemented in many different surveys. The generic approach implies a set of parameters to be set instead of writing code.

This article will give an insight into the theoretical base for SELEKT. In the next section SE is described, both generally and specifically for SELEKT. Section 3 includes methods for computation of anticipated values and intervals of normal variation from background data. Two major views on how to set the limit for the manual editing by SE are presented in Section 4. Finally, experience from implementation and running of SE at Statistics Sweden is summarised in Section 5.

2. Selective Editing

Suspicion of a data value being in error and the potential impact of a suspected error on output statistics are the two aspects to consider in the search for influential errors. SELEKT produces local scores based on indicators of both suspicion and impact for all variables, all statistical measures, and all domains of study that are considered important. These local scores are aggregated up to a global score for the respondent.

Example 1: Jäder and Norberg (2006) compute a score function for the International Trade in Goods statistics (ITG). Suspicion is based on observed price per quantity for a transaction, relative to normal variation in product groups, countries and direction (imports/exports). Potential impacts on estimates for domains of study are measured as the difference between reported value and anticipated value of trade, relative to the normal size of the domains. In this survey, suspicion and potential impact have a very low correlation and both of these dimensions are important.

Suspicion indicated by traditional edit rules is described in Subsection 2.1 along with alternative and supplementary methods to indicate suspicion. In Subsection 2.2, classes of domains of study in the output statistics are established and impact of errors is defined. In Subsection 2.3, suspicion and impact are combined and local scores are defined. These local scores are then aggregated to the global score in Subsection 2.4.

2.1. Suspicion

An edit, also known as an edit rule or a checking rule, can indicate that a data value is or might be in error. A fatal edit manifests if a data item is in error. Examples of fatal errors are inconsistent responses, invalid entries, and item nonresponse. Query edits point to suspicious data items. An example could be a value that, compared to historical data, seems suspiciously high. SELEKT makes use of indicators of the degree of suspicion, not only the traditional dichotomous results "accept" and "fail". Three options are available for assigning a level of suspicion with SELEKT: use of traditional edits, use of hit rates for edits, and the SELEKT-type edits. These are presented in Subsections 2.1.1, 2.1.2, and 2.1.3 respectively.

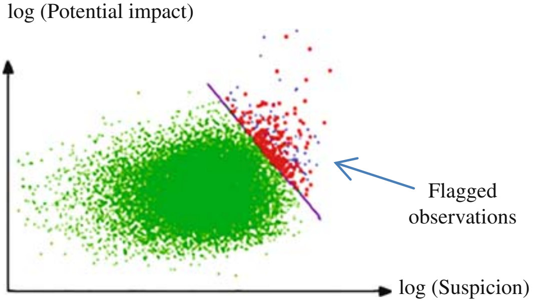


Fig. 1. International Trade in Goods statistics. Suspicion is based on price per quantity; potential impact reflects value of trade.

2.1.1. Traditional Edits

An edit is a logical condition or a restriction on a variable value or a group of variable values which must be met if the data is to be considered acceptable. In the absence of a conceptual model for edits, Norberg (2011) proposes the notion of a *test variable*, which is a function of one or more survey variables, possibly including register variables. The test variable can be the survey variable as is. A common test variable is the ratio of the survey variable to the corresponding variable in a previous survey round. Comparing received data to register information is another example. The value of the test variable is compared with an *acceptance region*, which is the range (or set) of acceptable values for the test variable. The comparison is often made for subgroups of data that are homogeneous with respect to the test variable so that the acceptance regions will be tight and thereby the edit will be efficient. These so-called *edit groups* are defined by using auxiliary variables such as stratification variables or variables used to define domains of study, but need not only be these two types.

Example 2: Say that the survey variables Number of employees and Turnover are measured for enterprises in different industries in a survey. The test variable Turnover/Number of employees is tested with different acceptance regions for three sets of industries. The name of the edit is A01.

Test variable: $z =$ Turnover / Number of employees
Edit groups:
- ◦ Industry = M (Manufacture)
- ◦ Industry = W (Wholesale trade)
- ◦ Industry = R (Retail trade)

Acceptance regions:
- ◦ [257€, 550€] for Industry = M
- ◦ [505€, 1082€] for Industry = W
- ◦ [227€, 632€] for Industry = R.

Edit rule A01 in the software code is:
if Industry = ‘M’ and not (257 < z < 550)
or Industry = ‘W’ and not (505 < z < 1082)
or Industry = ‘R’ and not (227 < z < 632)
then Errcode_A01 = ‘Fail’

A failed query edit like this does not necessarily imply that any of the observed, unedited values of the two survey variables are in error. In the example, it may well be the Industry code. The conceptual model, regardless of any lack of universality, is used to construct implicit edits by SELEKT, which is described in Subsection 2.1.3.

At Statistics Sweden, the implementations of SELEKT primarily make use of the existing traditional edits, simply for practical reasons. Suspicion is set to one if the edit is failed and to zero if the data is accepted.

2.1.2. Traditional Edits and the Use of Hit Rates

The dichotomisation of suspicion by 0 / 1 or “Accept”/“Fail” entails a loss of information compared to using varying levels of suspicion. For example, the level of suspicion by a traditional edit can be set equal to the hit rate of the edit, based on evaluations of previous survey rounds. UNECE (2000) defines:

"Hit rate is the success rate of an edit; the proportion of error flags that the edit generates which generate a change of data in the follow-up process".

Example 3 (cont. from Example 2): Assume that the hit rate for edit rule A01 is known from past survey rounds to be around 70%. Now add the software script: if Errcode_A01 = 'Fail' then Suspicion_A01 = 0.7

When a hit rate is less than 0.8, say, the original edit can be made manifold to utilise information.

Example 4 (cont. from Example 3): When the hit rate for edit A01 is known to be 70%, evaluate previous survey data to find out which wider acceptance regions would have yielded a hit rate of around 90%, say. Excluding the data outside this second acceptance region but keeping data that were flagged by the first acceptance region, the hit rate will be less than 0.7, say 0.6. Let us now make two edits A01a and A01b with the following script.

```
if Industry = 'M' and not ($257 < z < 550$)
or Industry = 'W' and not ($505 < z < 1082$)
or Industry = 'R' and not ($227 < z < 632$)
then Errcode_A01a = 'Fail'
if Industry = 'M' and not ($182 < z < 697$)
or Industry = 'W' and not ($360 < z < 1370$)
or Industry = 'R' and not ($152 < z < 948$)
then Errcode_A01b = 'Fail'
if Errcode_A01a = 'Fail' then Suspicion_A01 = 0.6
if Errcode_A01b = 'Fail' then Suspicion_A01 = 0.9
```

Varying the level of suspicion has a direct proportional effect on the score and thereby makes a difference in terms of the priority given to observations far out from the other observations.

Fatal errors could also be treated in SE, such that the suspicion is set to one. SELEKT has an option to send all fatal errors to follow-up or only those that have a major impact on output statistics. The rest can be imputed in a later stage of the production process.

2.1.3. SELEKT-Type Edits

SELEKT has a module which constructs implicit edits. Such edits can be used as a complement to or substitute for the traditional edits described above. The user specifies an edit by a test variable. A test variable is an existing variable or an expression of variables in the dataset to be edited, as for traditional edits. Edit groups are automatically defined by SELEKT according to a few specifications by the user. These groups are also used to compute anticipated values, see Section 3. Implicit acceptance regions are computed by SELEKT based on intervals of variation estimated from previous survey rounds.

Tukey's exploratory data analysis, EDA, is an approach to analysing data sets to summarise their main characteristics, often with visual methods. The box plot, based on quartiles of a distribution, has been a basis for the development of the SELEKT-type edits.

In this respect, SELEKT has similarities with the Hidiroglou-Berthelot method. Unlike the Italian Selemix, there is no assumption of any explicit distribution of the errors.

Let *k,l* identify observed unit (element) *l* belonging to primary sampled unit (cluster) *k*. Unit *k,l* implies data on two levels, say an enterprise delivers data for all employees or all products. One-stage sampling is just a special case with this notation. The notation is also valid for two-phase sampling.

Assume that quartiles and medians are preferred as the basis for the edits. Alternatives are presented at the end of this subsection. Each observed unit *k,l* belongs to one and only one edit group *g*. For a set of data from previous survey rounds, let

$\tilde{z}_{i,g}^{L} =$ lower quartile of the *i:th* edited test variable values for the edit group *g*
$\tilde{z}_{i,g}^{U} =$ upper quartile of the *i:th* edited test variable values for the edit group *g*
$\tilde{z}_{i,g}^{M} =$ median of the *i:th* edited test variable values for the edit group *g*.

Let
$z_{i,k,l} =$ unedited value of the *i:th* test variable in the current, unedited data.

The parameter $\kappa > 0$ defines the "gap" of the value range where suspicion shall be zero, that is, the acceptance region. A small κ yields suspicions larger than zero already at a small deviation from the anticipated value $\tilde{z}_{i,g}^{M}$.

The parameter $\tau \geq 0$ defines the regions where suspicion grows from 0 to 1. With $\tau = 0$ the suspicion equals 1 outside the acceptance region defined by κ.

Definition of suspicion $\xi_{i,k,l}$ by *i:th* test variable $z_{i,k,l}$:

If $z_{i,k,l} \leq \tilde{z}_{i,g}^{M} - (\kappa + \tau) \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right)$ then $\xi_{i,k,l} = 1$

if $\tilde{z}_{i,g}^{M} - (\kappa + \tau) \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right) < z_{i,k,l} < \tilde{z}_{i,g}^{M} - \kappa \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right)$ then

$$\xi_{i,k,l} = \frac{\tilde{z}_{i,g}^{M} - \kappa \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right) - z_{i,k,l}}{\tau \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right)}$$

if $\tilde{z}_{i,g}^{M} - \kappa \cdot \left(\tilde{z}_{i,g}^{M} - \tilde{z}_{i,g}^{L}\right) \leq z_{i,k,l} \leq \tilde{z}_{i,g}^{M} + \kappa \cdot \left(\tilde{z}_{i,g}^{U} - \tilde{z}_{i,g}^{M}\right)$ then $\xi_{i,k,l} = 0$

if $\tilde{z}_{i,g}^{M} + \kappa \cdot \left(\tilde{z}_{i,g}^{U} - \tilde{z}_{i,g}^{M}\right) < z_{i,k,l} < \tilde{z}_{i,g}^{M} + (\kappa + \tau) \cdot \left(\tilde{z}_{i,g}^{U} - \tilde{z}_{i,g}^{M}\right)$ then

$$\xi_{i,k,l} = \frac{z_{i,k,l} - \left(\tilde{z}_{i,g}^{M} + \kappa \cdot \left(\tilde{z}_{i,g}^{U} - \tilde{z}_{i,g}^{M}\right)\right)}{\tau \cdot \left(\tilde{z}_{\mathrm{i},g}^{U} - \tilde{z}_{i,g}^{M}\right)}$$

if $z_{i,k,l} \geq \tilde{z}_{i,g}^{M} + (\kappa + \tau) \cdot \left(\tilde{z}_{i,g}^{U} - \tilde{z}_{i,g}^{M}\right)$ then $\xi_{i,k,l} = 1$

When the lower quartile equals the median, the suspicion $\xi_{i,k,l}$ equals 0 for observed values lower than the median and analogously for the right side of the distribution. This could of course be considered to be a problem, so a manual intervention is recommended. SELEKT gives easy access to the files of quartiles.

The suspicion function is illustrated in Figure 2. Notice that this function considers a skewed population by an asymmetric acceptance region. Generally, a logarithmic transformation for skewed test variables is also recommended.

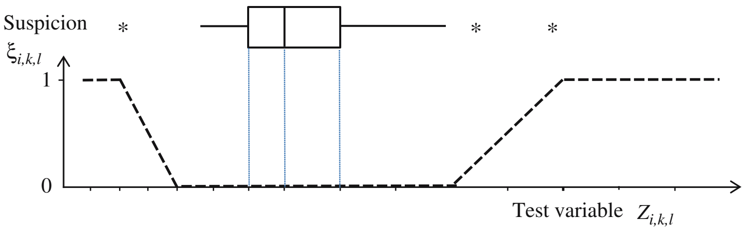


Fig. 2. Illustration of the SELEKT*-type method to measure suspicion. Assume that the test variable in historical edited data is distributed as shown by the box plot. With the parameters $\kappa = 3$ and $\tau = 2$, the suspicion function for new unedited data is demonstrated by the dashed line. Tick marks denote the distances (median – lower quartile) and (upper quartile – median), respectively.*

SELEKT has an option to use (arithmetic) mean – standard deviation, mean + standard deviation and the mean instead of the quartiles and the median. A new version of SELEKT is planned that will make use of time-series analysis with forecasts.

Suspicion for the unedited value of a survey variable j to be in error can be assigned by more than one edit. SELEKT simply uses $\xi_{j,k,l} =$ the maximum of the suspicions $\xi_{i,k,l}$ produced by all edits associated to survey variable j, both traditional and the SELEKT-type edits. This is good enough in many cases, but not always. If a variable value in the current survey round grossly fails an edit check compared with the previous month but is very well accepted in a check with the same month last year, the received data can be accepted, that is, the minimum of the two computed suspicions would be appropriate. In cases like this, a composite edit rule of "traditional type" must be specified.

2.2. *Impact*

This section begins by structuring the output statistics in Subsection 2.2.1. An erratic value in data can damage several estimated characteristics. In Subsection 2.2.2 the size of the error's impact on an estimate of a sum and on an estimate of a ratio are defined.

2.2.1. Produced Statistics – Output

Statistical information is built up from structured sets of estimates of statistical characteristics, that is, statistical tables (Sundgren 2001). An estimated statistical characteristic is defined as a statistical measure applied to the values of one or more variables in a set of objects, a domain of study. The statistical measure is an aggregating function, for example a function that counts the number of units, a function that summarises the values of a variable, and a function that computes the ratio between the sums of two variables. The domains of study are constructed by variables that cross classify the population.

For SELEKT to be comfortably generic, an agreed organisation of the domains of study and functions is required. The domains must be clustered into *classes* in such a way that there is no overlap between the domains within a class, that is, no observed unit may contribute to more than one domain in a class. It is described in Subsection 2.3 that classes can be assigned a different importance that applies to all domains of each class.

Let d denote a domain in a class c of domains.

Example 5: Say that there are five domains in a survey, all based on Industry code; M = Manufacture, T = Trade, subdivided into W = Wholesale trade and R = Retail trade and finally All = M + T. A "typical" table looks like this:

Table 1. Survey of employment and turnover (constructed data)

	Point estimates			Estimated standard errors		
Industry	Sum of number of employees	Sum of turnover	Turnover per employee	Sum of number of employees	Sum of turnover	Turnover per employee
M	470	172,358	367	11.2	5,171	11.0
T = W + R	473	246,345	521	9.5	12,317	20.9
W	197	141,950	721	0.0	0	0.0
R	276	104,395	379	13.8	9,396	30.3
All = M + T	943	418,703	444	10.6	13,359	9.1

The five domains can be clustered into three classes in two alternative ways. One clustering, seemingly attractive, is {All}, {M, T} and {W, R}. The other way is {All}, {M, W, R}, {T} where the second class is composed of domains at various levels of detail. A class of domains need not cover all units, as exemplified by the class {W, R} which does not cover employees in Manufacture.

2.2.2. Impact on Estimated Statistical Characteristics

Impact measures how much the unedited instead of the edited value would affect the output statistics. Let

$y_{j,k,l} =$ the unedited value of variable j and unit k,l
${}^{e}y_{j,k,l} =$ the edited value of variable j and unit k,l.

Estimate the sum of variable j in domain d in class c by summing the units in the sample that belong to the domain by:

$$\hat{T}_{c,d,j} = \sum w_{k,l} \cdot {}^{e}y_{j,k,l}$$

where $w_{k,l}$ is the combined sampling weight for the primary sampled unit k and observed unit l.

The impact on the estimate when retaining the unedited value for observed unit k,l belonging to c,d instead of the edited data value is

$$\eta_{c,d,j,k,l} = w_{k,l} \cdot \left(y_{j,k,l} -{}^{e} y_{j,k,l}\right)$$

In advance, ${}^{e}y_{j,k,l}$. is of course not available. As a proxy, an anticipated value $\tilde{y}^{M}_{j,k,l}$ is used, which can be the value from a previous survey round or an average of similar units in previous survey rounds. At this moment, it is not known whether the unedited value $y_{j,k,l}$ is in error or will be accepted. Potential impact is defined as:

$${}^{p}\eta_{c,d,j,k,l} = w_{k,l} \cdot \left(y_{j,k,l} - \tilde{y}^{M}_{j,k,l}\right)$$

For the ratio of the sums of variables *j1* and *j2,* the potential impact of *k,l* is:

$$
{}^{p}\eta_{c,d,j1,j2,k,l} = \frac{\hat{T}_{c,d,j1} + w_{k,l} \cdot \left(y_{j1,k,l} - \tilde{y}^{M}_{j1,k,l}\right)}{\hat{T}_{c,d,j2} + w_{k,l} \cdot \left(y_{j2,k,l} - \tilde{y}^{M}_{j2,k,l}\right)} - \frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}}
$$

$$
= w_{k,l} \cdot \frac{\left(y_{j1,k,l} - \tilde{y}^{M}_{j1,k,l}\right) - \dfrac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}} \cdot \left(y_{j2,k,l} - \tilde{y}^{M}_{j2,k,l}\right)}{\hat{T}_{c,d,j2} + w_{k,l} \cdot \left(y_{j2,k,l} - \tilde{y}^{M}_{j2,k,l}\right)}
$$

where $\hat{T}_{c,d,j1}$ and $\hat{T}_{c,d,j2}$ are estimated sums for variables *j1* and *j2* in domain *d* of class *c* from previous survey rounds. The ABS uses this potential impact as the local score according to Farwell (2004). A Taylor's approximation, which changes the denominator slightly if the anticipated error $y_{j2,k,l} - \tilde{y}^{M}_{j1,j2,g}$ is relatively small, is used on SELEKT:

$$
{}^{p}\eta_{c,d,j1,j2,k,l} = w_{k,l} \cdot \frac{\left(y_{j1,k,l} - \tilde{y}^{M}_{j1,\,k,l}\right) - \dfrac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}} \cdot \left(y_{j2,k,l} - \tilde{y}^{M}_{j2,k,l}\right)}{\hat{T}_{c,d,j2}}
$$

If the anticipated error is large, the potential impact will be large anyway.

2.3. *Local Scores*

Early approaches to SE had two steps. First the units were either accepted or flagged for having at least one variable value in error. Second, a potential impact of each suspected unit on produced statistics was estimated. Only the units with the most potential impact on the statistics would be prioritised for manual follow-up. The ABS tool for significance editing, SigEE, works according to this approach (Farwell 2004).

SELEKT calculates an anticipated impact per variable, measure and domain as the product of suspicion and potential impact. The local score for the observed unit *k,l* and the estimated sum of variable *j* for the domain *c,d* to which unit *k,l* is assumed to contribute is now defined as:

$$
SCORE5_{c,d,j,k,l} = \frac{\alpha_{c,d} \cdot \beta_j \cdot \left|\xi_{j,k,l} \cdot {}^{p}\eta_{c,d,j,k,l}\right|}{\max\left\{SE(\hat{T}_{c,d,j}),\, \delta \cdot \hat{T}_{c,d,j}\right\}^{\gamma_j}}
$$

where $SE(\hat{T}_{c,d,j})$ is the estimated standard error of the estimated sum of variable *j* in domain *c,d* from previous survey rounds, $\alpha_{c,d}$ is an importance weight for domain *c,d*, determined by subject-matter specialists and methodologists, β_j is a similar importance weight related to the estimated sum of variable *j*, also applied by Latouche and Berthelot (1992), δ and γ_j are parameters, $\delta > 0$ and $0 \leq \gamma_j \leq 1$.

This local score is the anticipated impact primarily relative to the standard error of the estimated sum. In sample surveys some domains may be totally enumerated, having zero standard errors. Some domains may have "accidentally" small estimated standard errors. SELEKT has one parameter δ to remedy these cases. A rather small value, like a general coefficient of variation in the survey, can be used as the value of δ. If sum is preferred to the standard error as the denominator, set δ to a big enough value, say 100.

SELEKT offers an option for assigning more importance to large domains and less importance to small domains, within the same class of domains. The denominator of the *SCORE*5 is raised to a power γ_j. The default value is 1. The value 0 reduces the denominator to a constant and anticipated impact is not related to any indicator of size of the domain. A "value in the middle" could be justified to obtain better relative quality of a large rather than a small domain, for example for "cars" rather than for "motorcycles" in the International Trade in Goods statistics.

In SE, as in survey design in general, it is necessary that the national statistical institute can assess the quality demands on each output table from the client's point of view.

The local score for two variables in a ratio estimated for a domain *c,d* is defined as

$$SCORE5_{c,d,j1,j2,k,l} = \frac{\alpha_{c,d} \cdot \beta_{j1,j2} \cdot \left| \xi_{j,k,l} \cdot {}^{p}\eta_{c,d,j1,j2,k,l} \right|}{\max\left\{ SE\left(\frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}} \right), \delta \cdot \frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}} \right\}^{\gamma_{j1,j2}}}$$

where $SE\left(\frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}}\right)$ is the estimated standard error of the estimated ratio and $\beta_{j1,j2}$ is the importance factor for the ratio of variables *j1* and *j2*.

SELEKT produces a "weight matrix" with all the domains *d* as rows, grouped by classes *c*. Variables for which sums and ratios of sums are estimated are represented as columns. The cells of this weight matrix, denoted C, are $\alpha_{c,d} \cdot \beta_j / \max\{SE(\hat{T}_{c,d,j}), \delta \cdot \hat{T}_{c,d,j}\}^{\gamma_j}$ for a sum and $\alpha_{c,d} \cdot \beta_{j1,j2} / \max\left\{SE\left(\frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}}\right), \delta \cdot \frac{\hat{T}_{c,d,j1}}{\hat{T}_{c,d,j2}}\right\}^{\gamma_{j1,j2}}$ for a ratio.

The domain importance weights $\alpha_{c,\,d}$ are initially equal for all domains belonging to a class *c*. Once the matrix is calculated, the product manager is free to change any value in the matrix, for any reason.

Example 6 (cont. from Example 5): Number of employees and Turnover are measured for enterprises in a survey. Estimates are produced for five domains. The five domains are grouped into three classes in such a way that there is no overlap of domains within a class. Let us set the following parameter values: $\delta = 0.02$.

Table 2. Importance weights for classes of domains

Class of domains *(c)*	$\alpha_{c,\,d}$
$c = 1$: M, T	1
$c = 2$: W and R	1
$c = 3$: M + T	0.5

Table 3. Importance weights for variables and measures

	Variables/measures		
	Sum of Number of employees $j = 1$	Sum of Turnover $j = 2$	Turnover per employee $j1 = 2$ & $j2 = 1$
β_j, $\beta_{j1,j2}$	1	1	5
γ_j, $\gamma_{j1,j2}$	0.5	0.5	1

Now the resulting weight matrix C is computed (by SELEKT):

Table 4. The weight matrix C (result)

Class and domain *(c,d)*	Sum of number of employees	Sum of turnover	Turnover per employee	Examples of computation
1, 1: M	0.299	0.0139	0.455	$1 \cdot 1/\max\{11.2,\ 0.02 \cdot 470\}^{0.5} = 0.299$
1, 2: T	0.325	0.0090	0.240	$1 \cdot 1/\ \max\{9.5,\ 0.02 \cdot 473\}^{0.5} = 0.325$
2, 3: W	0.504	0.0188	0.347	$1 \cdot 1/\ \max\{0,\ 0.02 \cdot 197\}^{0.5} = 0.0188$
2, 4: R	0.269	0.0103	0.165	$1 \cdot 1/\max\{9\ 396,\ 0.02 \cdot 104\ 395\}^{0.5}$ $= 0.0103$
3, 5: M + T	0.115	0.0043	0.278	$0.5 \cdot 5/\ \max\{9.1,\ 0.02 \cdot 444\}^{1} = 0.278$

When the matrix is computed by SELEKT, it is available and it is possible to alter any value, either ad hoc or by using an alternative model of computation.

2.4. Aggregation of Local Scores to Global Scores

There is a hierarchy of scores with five levels. Let r denote respondent. In most surveys, the respondent answers for only one primary sampled unit, but there are examples where the respondent has many sampled units. The levels are:

5: Domains (d),
4: Variables/measures (j or $j1, j2$),
3: Observed units (l)
2: Primary sampled units (k)
1: Respondent (r).

The numbering from 5 to 1 makes space for additional levels in future versions, if needed. Additional levels conceivably could be survey round and edit rule.

Scores for respondents, if being calculated, are global scores. Most often the scores for primary sampled units are considered the global scores. If the cost for recontacts in the follow-up workload is negligible, the scores for observed units can be used as global scores.

The local score for variable/measure j or $j1,j2$ is an aggregate of scores defined above:

$$SCORE4_{j,k,l} = \left\{ \sum_{c} \left[max\left(0, SCORE5_{c,d,j,k,l} - \theta_5\right)\right]^{\lambda_5} \right\}^{1/\lambda_5} \quad \text{or}$$

$$SCORE4_{j1,j2,k,l} = \left\{ \sum_{c} \left[max\left(0, SCORE5_{c,d,j1,j2,k,l} - \theta_5\right)\right]^{\lambda_5} \right\}^{1/\lambda_5}$$

where the summing $\sum_c$ is over all classes of domains (or domains since each selected unit contributes to one or no domain in each class of domains).

θ_5 is a threshold parameter. λ_5 is a parameter that defines the aggregation method. Hedlin (2008) presents the simple function used here to distinguish between three options. The value 1 implies sum, the value 2 implies sum of squares and the value 100 implies maximum (in practice). Latouche and Berthelot (1992) suggest the sum of local scores for variables, whereas Lawrence and McDavitt (1994) and Hedlin (2003) use the maximum of the local scores. Farwell (2005) proposes the Euclidian distance, that is the value 2, as an alternative to the maximum when there are many scores at the variable level. SELEKT has four parameters; λ_5, λ_4, λ_3 and λ_2, one for each level to be aggregated.

The aggregation of local scores continues step by step up to the respondent level as:

$$SCORE3_{k,l} = \left\{ \sum_j \left[max\left(0, SCORE4_{j,k,l} - \theta_4\right)\right]^{\lambda_4} \right\}^{1/\lambda_4}$$

$$SCORE2_k = \left\{ \sum_l \left[max\left(0, SCORE3_{k,l} - \theta_3\right)\right]^{\lambda_3} \right\}^{1/\lambda_3}$$

$$SCORE1_r = \left\{ \sum_k \left[max\left(0, SCORE2_k - \theta_2\right)\right]^{\lambda_2} \right\}^{1/\lambda_2}$$

At all aggregations, SELEKT allows for a threshold value so that the maximum of zero and the score minus this threshold is aggregated. A value > 0 in early aggregation steps can be justified if there is a marginal cost of manual follow-up of extra variables, for example, with the constraint that the respondent has already been contacted. There are four threshold parameters θ_5, θ_4, θ_3 and θ_2. θ_1 is the threshold value for the global score, see Subsection 4.1.

$SCORE2_k$ is the global score in most surveys. An aggregation of scores for primary sampled units to $SCORE1_r$, where r primarily denotes respondent, is of interest in special cases. At Statistics Sweden the following situations have occurred:

- Respondents have several primary sampled units in the sample in the annual survey Rents for Dwellings. The national sample consists of approximately 12,000 rented dwellings. It turns out that there are only about 2,600 respondents, the real estate owners. The selective editing is focused on minimising the number of recontacts to respondents.
- The inflow of data for monthly surveys is irregular in the sense that two or more monthly forms can be delivered in the same "batch". Month is not defined as a level in the hierarchy of data. The $SCORE2_k$ would be an aggregate of scores for all months unless any action is taken. By defining Level 1 as the combination of primary sampled unit and month, that is a more detailed level than Level 2, the $SCORE1_r$ returns scores per primary selected unit and month.

3. Anticipated Value and a Measure of Variation

This section concerns methods and data to use for the computation of anticipated values and a basis for the SELEKT-type edits.

3.1. Time-Series Versus Cross-Section Data

There are quite different approaches to find anticipated values for the survey variables (j) and the test variables (i) and intervals of variation for the test variables for each observed unit s (k,l):

- Forecasts from the analysis of time-series data per observed unit.
- Latest observed values from previous survey rounds. Intervals of variation for the test variables cannot be computed.
- Cross-sectional analysis of data from previous survey rounds. Firstly find homogeneous groups of units, secondly compute the anticipated value as the median, the arithmetic mean or any other central value in the homogenous groups and also intervals of variation.

Intuitively, it makes sense to set the time-series measures as priority and the cross-sectional measures as reserve. Most business survey designs include updates of samples annually or more often. With a scheme of rotating samples, quite a significant proportion of units are new in the sample each year. This implies that different methods need to be used for different units if one wants to use the best method when possible.

3.2. Cold-Deck and Hot-Deck Data

The SE needs anticipated values and measures of variation before the editing of a survey round to make editing of data possible as soon as the first data arrive. Calculations of these measures are often based on edited data from past survey rounds, so-called cold-deck data. Generally calculations are made without using sampling weights. A decision has to be made whether to include imputed data. It seems most advisable from a theoretical point of view not to use imputed data, but it is easier not to distinguish between imputed and collected data. A decision must also be made about whether to include data that were suspicious but not flagged because the potential impact was low. Again it seems to be a good idea not to use highly suspected data, but it is easier not to make a difference.

For advanced time-series models, at least three years of monthly/quarterly data are needed. The simplest method in this context is to use the latest value or a simple function of it as anticipated value. If the survey measures phenomena with a heavy seasonal pattern, such as turnover in retail trade, a better alternative can be the value of the same month last year.

For cross-sectional data, one survey round can suffice, although several are preferable. Data from the current survey round, so-called hot-deck data, can be used by successively updating anticipated values and variation. The editing of prices for fresh fruits and vegetables in the consumer price index preferably makes use of robust means of current prices and price ratios. Prices from last month may be obsolete due to rapid price changes for fresh products.

3.3. Homogeneous Groups for Cross-Sectional Analysis

Cross-sectional analyses require homogeneous groups for which anticipated values are computed. The groups may, but need not, correspond to strata or domains of study. It is more important to stress homogeneity rather than to have a large number of observations in the homogeneous group. A well-known technique from the literature for imputation is to use the value of the very nearest neighbour as the anticipated value. However, by doing so, there is only one observation and it is not possible to compute an interval of variation for the SELEKT-type edits.

Homogeneous groups should be defined by multivariate analysis of cold-deck data. There are various methods that can be used. Norberg (2012) demonstrates regression tree analysis. The result of such an exercise is a new variable that identifies the groups.

SELEKT has one module that constructs joint homogeneous groups for the purposes of computing anticipated values and edit groups for SELEKT-type edits. Classificatory variables are listed by the user in hierarchical order and the minimum number of observations needed in the groups is specified. The cold-deck data are successively split by the classificatory variables, one by one, into one group for each value in the value set of the classificatory variable, as long as the condition for minimum number of observations is satisfied. The groups can be defined not only by different variables, but also by different numbers of digits within the classification codes.

3.4. *Anticipated Value and Variation*

For time-series data, there are a few options. The simplest is to use the edited values from the latest survey round for unit *k,l*. Here no measure of variation of the test variable can be computed. Another option is to use observations across a number of previous survey rounds and choose either the median, the lower and upper quartiles or arithmetic mean ± standard deviation. A third option is to produce a forecast for unit *k,l* by performing a time-series analysis including confidence intervals for the forecast.

For cross-sectional analysis, there are two natural sets of measures within the homogeneous groups; firstly, the lower quartile, median and upper quartile computed unweighted across the previous survey rounds for homogeneous groups, and secondly, the arithmetic mean and standard deviation.

SELEKT allows for auxiliary variable values $x_{j,k,l}$ which might help to compute the anticipated value $\tilde{y}^M_{j,k,l}$ of the *j:th* variable. If the ratios of $y_{j,k,l}/x_{j,k,l}$ have a small variation, the anticipated value should firstly be found for the ratio and secondly this anticipated ratio should be multiplied by the individual $x_{j,k,l}$ to yield an anticipated value of $\tilde{y}^M_{j,k,l}$. The auxiliary variable should preferably be almost error free, so as not to cause many high scores by itself. The International Trade in Goods statistics are a converse example; the anticipated invoiced value of a transaction is the observed quantity multiplied by the median of price per quantity for transactions in the homogeneous group. There are more often errors in the quantity than in the invoiced value, but no better anticipated value can be found for invoiced value.

Anticipated values can be estimated by regression analysis with several explanatory variables. It is not a major problem to add files of anticipated values from tailor-made analysis to the SELEKT system.

As already noted in Subsection 3.1, one must be prepared to use different methods for computing the anticipated values and measures of variation. Different methods might be used for variables but particularly for various parts of the data. The SE can be inefficient if the “precision” of the anticipated values and intervals of variation varies. In Subsection 2.1.2 the indicator hit rate was defined. It is straightforward to compute this indicator for new units in the sample and those with a long series of data. Anticipated values, used in the impact dimension, can be analysed with edited values in scatter plots. The scores (being anticipated impacts relative to the standard error of the estimated sum) include both

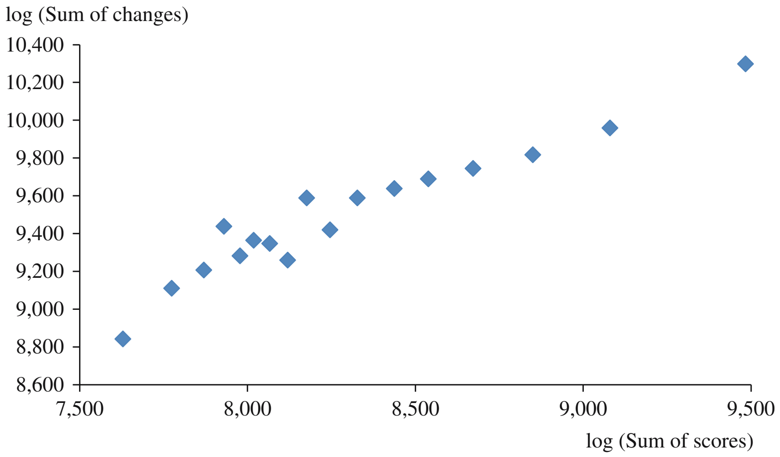


Fig. 3. International Trade in Goods statistics. Vertical axis is sum of absolute values of changes on invoiced value. Horizontal axis is sum of scores. Each point represents an aggregate of 50 observed units for 70 months, in order of score. The very highest scores are far out and are not included in the graph.

dimensions. The graph in Figure 3 is a means to see if identified errors on average are proportional to scores for a survey with one prioritised variable, the sum of which is estimated. Some flagged units did not result in a change, some did. Batches of observations along the scale of score make computations of sum of impacts possible. As the scores in Figure 3 are fairly proportional to the changes on average, they could be considered useful predictors in the search for frequent and/or big errors. With this technique, it would also be possible to analyse any differences due to methods for estimation of anticipated values and so on by plotting two or more series of data. Surveys with many variables, many classes of domains and varying importance weights get a complex global score that it is scarcely possible to analyse this way.

4. How Much is Enough?

This section presents two different ways to use the scores to identify data for follow-up. The traditional selective data editing approach selects units with scores above a threshold, called cut-off selection, as is discussed in Subsection 4.1. In Subsection 4.2 a new idea suggesting probabilistic editing is presented briefly. This second method is not yet implemented at Statistics Sweden.

4.1. Cut-Off Selection

To determine cut-off thresholds for the global score, many calculations have to be carried out before the implementation using data from earlier survey rounds. Edited and unedited data are required. The concept of pseudobias is easy to understand and use.

Latouche and Berthelot (1992) define absolute pseudobias for an estimate when Q percent of the primary selected units with the highest scores are followed up as

$$\frac{\hat{T}_{c,d,j,Q} - \hat{T}_{c,d,j,Q=100}}{\hat{T}_{c,d,j,Q=100}}$$

where $\hat{T}_{c,d,j,Q}$ is an estimate of the sum of variable j in a domain c,d. $\hat{T}_{c,d,j,Q=100}$ is the estimated sum when all data that have been followed up are included (approximated by the old traditional heavy editing).

Lawrence and McDavitt (1994) define relative pseudobias, *RPB*, as

$$RPB_{d(s),j,Q} = \frac{\hat{T}_{c,d,j,Q} - \hat{T}_{c,d,j,Q=100}}{SE(\hat{T}_{c,d,j,Q=100})}$$

where $SE(\hat{T}_{c,d,j,Q=100})$ is the estimated standard error of $\hat{T}_{c,d,j,Q=100}$. Whenever $SE(\hat{T})$ in the denominator is close to zero, it should be replaced with some fraction of $\hat{T}$ as for SCORE5 in Subsection 2.3.

Särndal et al. (1992, 165) show that a 20% *RPB* has little effect on the coverage probability of an estimated confidence interval based on the estimated sampling variance. Allowing various importance parameters for (classes of) domains and variables/measures creates complications. It will not be as simple as requiring that all *RPBs* be less than 20%, but primarily the *RPBs* of the most important output should be less than 20%, while the rest can do with a somewhat higher *RPB*. Furthermore, the randomness of measurement errors in an evaluation data set occasionally causes high *RPB* in some domains. Evaluation as such should thus be done on more than one survey round. Hence, requiring *RPBs* less than 20% is a rule of thumb for important variables and domains, while accepting a few higher *RPBs* for less important statistical characteristics. Small domains can get a high *RPB* by accident even though errors are completely at random. When systematic errors frequently exist in data it will most likely be difficult to find a cut-off threshold that reduces the editing workload. MEMOBUST (2014) has three remarks on the method:

- The assumption of this approach is that the edited data can be considered 'true' data. This is a limitation because it rarely can be assumed.
- The simulation approach is frequently applied to data of a previous survey occasion to obtain a threshold value to be used for the current survey. It is worthwhile to note that in this case we assume that the error mechanism and the data distribution are the same on the two occasions.
- The method cannot be applied when you deal with the first wave of a survey.

For smaller recurrent surveys, such as industrial production, a feasible approach is to allow a real-time assessment by professional editing clerks based on the score itself, which may or may not be complemented by a ranking variable.

As a supplement to cut-off selection of units to follow up, a random sample beyond the threshold is useful for evaluating the performance of the SE in the long run. The results of the sample will indicate when the thresholds need to be updated with current data. Lewis (2014) discusses the need to formally maintain selective editing systems used in business surveys. When selective editing is first introduced, it offers the opportunity for efficient micro editing. However, without regular review, the thresholds can become out of date, potentially leading to an inefficient process and low-quality outputs.

4.2. Probabilistic Editing

Ilves and Laitila (2009) and Ilves (2010) propose quite a different editing procedure, where the responses are selected for editing through Poisson sampling according to their anticipated impact on final estimates, that is, the global scores. The probabilistic approach uses simple tools known from sampling theory to describe the effect of editing on the survey estimates. There is no need to restrict the follow-up to suspected data above a threshold. A two-phase design approach is applied to the bias estimation and a bias-corrected generalised regression (GREG) estimator and its variance are presented. Since the impact of systematic measurement errors is possible to estimate, it is not as important as for the cut-off method to identify and fix these errors before the selective editing.

5. Experience

Statistics Sweden has implemented selective editing in eleven surveys with extensive data editing over the last years and further surveys are in the pipeline for implementation (Norberg et al. 2014).

The experience is that implementation is a resource-intensive task that includes:

- staging tables of microdata and output statistics from SQL databases,
- creating performance indicators of the existing traditional edits,
- finding homogeneous edit groups by multivariate analysis,
- finding threshold values for local and global scores,
- integrating the SAS®-based SELEKT with production systems programmed in VB6 or VB.Net.

Efficient edit rules, as the result of reviewing the existing edits, are a basis for good data quality. The implementation of SE should at an early stage render in an evaluation of the existing traditional edit rules based on some hit-rate indicator. More efficient edit rules have been implemented as a result of these reviews.

It is necessary to address the question “Is selective editing appropriate for the survey?” as early as possible in the implementation stage. Statistics Sweden has developed a checklist based on experience. The checklist contains considerations of the following aspects:

- key variables are continuous,
- systematic measurement errors for known causes are dealt with in a first step,
- outputs are aggregates of microdata (statistical characteristics),
- anticipated values are possible to obtain.

The best profits were achieved for surveys where

- microediting is extensive, there is a potential for savings,
- the survey design is such that units have very different impacts on the estimates,
- there is a limited number of important variables and classes of domains, otherwise it will not be possible to define the parameters for the score function and global threshold so that the relative pseudobias is less than 20% percent for most statistical estimates.

The selective editing at Statistics Sweden has resulted in:

- reduction of error lists by 10–60% and consequently a reduction of cost for running production,
- a replacement of late macroediting by early microediting which implies that follow-ups can be done closer to data capture which is beneficent for respondents and measurement quality,
- more effective, more interesting and less stressful work for the editing staff,
- reduction of mishaps of records slipping through the editing since the editing staff now receive shorter error lists with a priority stated for all units.

6. References

Adolfsson, C. and P. Gidlund. 2008. "Conducted Case Studies at Statistics Sweden." Paper presented at the Work Session on Statistical Data Editing, Vienna, Austria, 21–23 April 2008. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.32.e.pdf (accessed February 2016).

Brinkley, E., K. Farwell, and F. Yu. 2011. "Selective Editing Methods and Tools: An Australian Bureau of Statistics Perspective." In Proceedings of Statistics Canada Symposium 2011. Available at: http://publications.gc.ca/collection_2013/statcan/11-522-x/CS11-522-2011-eng.pdf (accessed February 2016).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. New York: Wiley.

Di Zio, M. and U. Guarnera. 2013. "A Contamination Model for Selective Editing." *Journal of Official Statistics* 29: 539–555. Doi: http://dx.doi.org/10.2478/jos-2013-0039.

Farwell, K. and M. Raine. 2000. "Some Current Approaches to Editing in the ABS." In Proceedings ICES II of the Second International Conference on Establishment Surveys, Invited Papers. American Statistical Association 2001, 529–538.

Farwell, K. 2004. "The General Application of Significance Editing to Economic Collections." Australian Bureau of Statistics. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.260.1420 (accessed February 2016).

Farwell, K. 2005. "Significance Editing for a Variety of Survey Situations." Paper presented at the 55th Session of the International Statistical Institute, Sydney, 5–12 April 2005.

Granquist, L. 1984. "On the Role of Editing." *Statistical Review* 2: 105–118.

Granquist, L. 1997. "The New View on Editing." *International Statistical Review* 3: 381–387. Doi: http://dx.doi.org/10.2307/1403378.

Granquist, L. and J. Kovar. 1997. "Editing of Survey Data: How Much Is Enough?" *Survey Measurement and Process Quality*, 415–435. Doi: http://dx.doi.org/10.1002/9781118490013.ch18.

Hedlin, D. 2003. "Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics." *Journal of Official Statistics* 19: 177–199.

Hedlin, D. 2008. "Local and Global Score Functions in Selective Editing." Paper presented at Work Session on Statistical Data Editing, Vienna, 21–23 April 2008.

Available at: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.31.e.pdf (accessed February 2016).

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

Ilves, M. and T. Laitila. 2009. "Probability-Sampling Approach to Editing." *Austrian Journal of Statistics* 38: 171–182. Available at: http://www.stat.tugraz.at/AJS/ausg093/093Ilves.pdf (accessed February 2016).

Ilves, M. 2010. "Probabilistic Approach to Editing." Workshop on Survey Sampling Theory and Methodology, Vilnius, Lithuania, August 23–27, 2010. Available at: https://www.amstat.org/sections/srms/proceedings/y2010/Files/308253_60434.pdf (accessed February 2016).

Jäder, A. and A. Norberg. 2006. "A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics." Background facts on Economic Statistics 2006:3, Statistics Sweden. Available at: http://www.scb.se/statistik/_publikationer/OV9999_2006AOLBR_X100ST0603.pdf (accessed January 2016).

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

Lawrence, D. and C. McDavitt. 1994. "Significance Editing in the Australian Survey of Average Weekly Earnings." *Journal of Official Statistics* 10: 437–447.

Lawrence, D. and R. McKenzie. 2000. "The General Application of Significance Editing." *Journal of Official Statistics* 16: 243–253.

Lewis, D. 2014. "Maintenance of Selective Editing in ONS Business Surveys." Paper presented at Work Session on Statistical Data Editing, Paris, France, 28–30 April 2014. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_1_ONS_Lewis.pdf (accessed February 2016).

MEMOBUST. 2014. *Handbook on Methodology of Modern Business Statistics CROS-portal*, Eurostat. Available at: https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_eng (accessed February 2016).

Norberg, A. 2011. "The Edit." Paper presented at Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9–11 May 2011. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.26.e.pdf (accessed February 2016).

Norberg, A. 2012. "Tree Analysis – A Method for Constructing Edit Groups." Paper presented at Work Session on Statistical Data Editing, Oslo, Norway, 24–26 September 2012. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/05_Sweden.pdf (accessed February 2016).

Norberg, A., K. Lindgren, and C. Tongur. 2014. "Experiences from Selective Editing at Statistics Sweden." Paper presented at Work Session on Statistical Data Editing, Paris, France, 28–30 April 2014. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_1_Sweden_Norberg.pdf (accessed February 2016).

Sundgren, B. 2001. "The αβγτ-Model: A Theory of Multidimensional Structures of Statistics." Paper prepared for the MetaNet conference in Voorburg, the Netherlands, 2–4 April 2001. Available at: https://www.google.se/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjb44jljfLKAhVjZ3IKHVUoC44QF

ggjMAA&url=https%3A%2F%2Fsites.google.com%2Fsite%2Fbosundgren%2Fmy-life%2FAlfaBetaGammaTauFinal.doc%3Fattredirects%3D0&usg=AFQjCNFKLgmtI-LR9YrLsQglsNWXnUY2mA&bvm=bv.114195076,d.bGQ (accessed February 2016).

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

UNECE. 2000. *Glossary of Terms on Statistical Data Editing*. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/editingglossary.pdf (accessed February 2016).

Received February 2013
Revised October 2015
Accepted December 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 231–256, http://dx.doi.org/10.1515/JOS-2016-0011

Synthetic Multiple-Imputation Procedure for Multistage Complex Samples

Hanzhi Zhou[1], *Michael R. Elliott*[2], *and Trivellore E. Raghunathan*[3]

Multiple imputation (MI) is commonly used when item-level missing data are present. However, MI requires that survey design information be built into the imputation models. For multistage stratified clustered designs, this requires dummy variables to represent strata as well as primary sampling units (PSUs) nested within each stratum in the imputation model. Such a modeling strategy is not only operationally burdensome but also inferentially inefficient when there are many strata in the sample design. Complexity only increases when sampling weights need to be modeled. This article develops a general-purpose analytic strategy for population inference from complex sample designs with item-level missingness. In a simulation study, the proposed procedures demonstrate efficient estimation and good coverage properties. We also consider an application to accommodate missing body mass index (BMI) data in the analysis of BMI percentiles using National Health and Nutrition Examination Survey (NHANES) III data. We argue that the proposed methods offer an easy-to-implement solution to problems that are not well-handled by current MI techniques. Note that, while the proposed method borrows from the MI framework to develop its inferential methods, it is *not* designed as an alternative strategy to release multiply imputed datasets for complex sample design data, but rather as an analytic strategy in and of itself.

Key words: Finite population Bayesian bootstrap; Haldane prior; stratified sample; clustered sample; sample weights.

1. Introduction

Stratified multistage sampling is the most common type of sample design for large-scale surveys conducted by the U.S. federal statistical agencies. This type of sample design combines the advantages of both stratification (for statistical efficiency) and cluster sampling (for cost and logistical efficiency). Under this design, the primary sampling units (PSUs) are stratified in such a way that they are homogeneous with respect to a stratum-level aggregate of the variable(s) of interest. To permit a maximum degree of stratification

[1] Mathematica Policy Research, Princeton, NJ 08543, USA. Email: zhouhanz@umich.edu.
[2] Dept. of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI USA 48109; Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48109, USA. Email: mrelliot@umich.edu.
[3] Dept. of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI USA 48109; Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48109, USA. Email: teraghu@umich.edu.

Acknowledgments: This work was supported in part by Grant Number R01CA129101 from the National Cancer Institute. The authors would like to thank Rod Little, Brady West, and Richard Valliant, along with the Associate Editor and three reviewers, for their review and helpful comments.

and thus variance reduction, it is common practice to define a large number of strata where only a small number of PSUs are selected in each stratum.

Multiple imputation (MI) (Rubin 1976, 1987) is a method commonly used when item-level missing data are present. However, MI requires that survey design information be built into the imputation models. Reiter et al. (2006) demonstrated the importance of simultaneously accounting for stratum effects and clustering effects in multiple imputation. They showed that when design features were ignored in the imputation model, biases would occur on the estimated parameter, even if a design-based analysis method was applied to the imputed data. Current MI methods typically include dummy variables to represent strata as well as PSUs nested within each stratum in the imputation model. When necessary, they also identify statistically significant interactions between these dummies with other covariates through routine variable selection procedures such as stepwise regression (Reiter et al. 2006; Schenker et al. 2006). Such a modeling strategy is not only operationally burdensome but also inferentially inefficient when there are hundreds of strata in the sample design and the sample in each stratum consequently becomes sparse. For example, the Census Bureau's Current Population Survey design groups 1,768 nonself-representing PSUs into 220 strata.

Possibly a better strategy is to consider clusters as random effects while treating strata as either fixed (using dummies) or random effects. However, many of the popular software packages that implement multiple imputation (e.g., SAS MI procedure, R packages *mice* or *mi*, and IVEware) cannot be adapted easily to such an approach. While a few recent software modules (such as R package *pan* and MLwiN module *REALCOM-IMPUTE*) have started to incorporate mixed effects or multilevel modeling for imputation purposes, they typically assume normal or latent normal distribution for variables with missing data. Their performances for missing categorical variables (binary in particular) are unclear. Moreover, there has been only little research that formally investigates their use in incorporating strata as well as clusters.

To circumvent these problems with fully parametric model-based imputation techniques, we develop a two-step semiparametric MI method. The idea is to separate the need to account for complex sample designs from the treatment of missing data. In the first step, sample designs are "reversed" through synthetic population data generation using a weighted finite population Bayesian bootstrap (FPBB) (Cohen 1997; Little and Zheng 2007; Dong et al. 2014). In the second step, missing values are imputed in the created synthetic population based on a parametric model that assumes identically independently distributed (IID) data elements. To account for stratum effects, we combine a replication variance estimation method (Efron 1979; Kovar et al. 1988; Rao and Wu 1988; Rao et al.1992; Rust and Rao 1996) with the weighted FPBB. Under a standard missing at random (MAR) assumption (Little and Rubin 2002), our method requires neither complicated modeling of strata and clusters nor design-based analyses of the imputed data. Note that while the proposed method borrows from the multiple-imputation framework to develop its inferential methods, it is *not* designed as an alternative strategy to release multiply imputed datasets for complex sample design data. Rather, it is intended an alternative analytic strategy for population inference from complex sample design data with item-level missingness.

In this article, we focus on the estimation of two quantities: quantile estimates for a continuous variable, and estimates of rare proportions and their associated logistic regression estimates. We consider a stratified two-stage sample design and investigate a full range of quantiles including tail behaviors. While design-based methods for quantile estimation from complex survey data have been developed (Francisco and Fuller 1991; Woodruff 1952), quantile estimation after imputation is less commonly addressed in the literature. (A recent exception that considers nonparametric fractional imputation outside of the complex sample design setting is Yang et al. 2013.) This is the case despite the rapid development and increasing popularity of MI. We also consider MI for incomplete binary variables, with a focus on rare outcomes. It is well known that maximum-likelihood estimation of logistic regression models typically suffers from small sample bias, the degree of which is strongly dependent on the number of sample cases in the less frequent of the two categories (King and Zeng 2001). Thus when the dependent binary variable represents the occurrence of rare events, the logistic regression coefficients can be substantially biased even with a simple IID data structure. Random effects logistic models are commonly used for fitting clustered binary data; however, these models rely heavily on asymptotic theory assumptions, which may not be met in sparse samples. All these issues might extend naturally to the missing-data context. As shown by Zhao and Yucel (2009), sequential MI for binary data missing completely at random in a multilevel setting suffers from severe bias and poor coverage in estimating probabilities that are close to 0 or 1, particularly when the intraclass correlation is high.

The objectives of this article are: i) to develop a two-step synthetic MI method as a way to simultaneously account for stratification, clustering, and unequal inclusion probability; and ii) to demonstrate the effectiveness of the new method with respect to quantile estimation and logistic regression for binary rare events data as compared with existing fully parametric imputation strategies. Section 2 discusses the imputation strategies under three different models: simple random sample, fixed effects for clusters/strata, and random effects for cluster/strata. Section 3 introduces the newly proposed procedure and the MI inference rules for quantile estimation under this method. Section 4 presents a Monte Carlo simulation study as the validation tool to assess the repeated sampling properties of MI under the various approaches. Section 5 applies different MI procedures to the analysis of body mass index on youth data from the third National Health and Nutrition Examination Survey (NHANES III). Some concluding remarks follow in Section 6. We focus on the two-PSU-per-stratum design in this chapter, although the methods we develop can accommodate any number of PSUs per stratum.

1.1. Fully Parametric Imputation Methods for the Two-PSU-per-Stratum Design

Here we briefly describe fully parametric multiple-imputation techniques with complex sample design features incorporated to different degrees. We assume the missing data Y_i is a member of the exponential family, and that there are fully observed covariates X_i (a $(p+1)$-dimension vector) such that $g(E(Y_i|X_i)) = X_i\beta$ for a known link function $g(\cdot)$ (e.g., $g(u) = \log(u/(1-u))$ for binary outcomes (logistic regression), $g(u) = \log(u)$ for count outcomes (Poisson regression), or $g(u) = u$ for continuous outcomes (Gaussian regression)).

1.1.1. Standard Regression Model Assuming SRS

Based on the maximum-likelihood estimates $\hat{\beta}$ and the associated asymptotic covariance matrix $\hat{V}(\hat{\beta})$ for the generalized linear model $g(E(Y_i|X_i)) = X_i\beta$, the posterior predictive distribution of the parameters can be constructed, which is then used to impute the missing values (Rubin 1987, 169–170). Point and variance estimates of the regression parameters can then be obtained using the usual MI combining rules (Rubin 1987, 76). For the p^{th} component of the regression parameter:

$$\hat{\beta}_p = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_p^{(m)}, \tag{1}$$

$$\hat{V}(\hat{\beta}_p) = \frac{1}{M}\sum_{m=1}^{M}\hat{V}\left(\hat{\beta}_p^{(m)}\right) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_p^{(m)} - \hat{\beta}_p\right)^2 \tag{2}$$

and

$$\frac{(\hat{\beta}_p - \beta_p)}{\sqrt{\hat{V}(\hat{\beta}_p)}} \dot{\sim} t_\nu,\ \nu = (M-1)\left(1 + \frac{\sum_{m=1}^{M}\hat{\beta}_p^{(m)}}{\frac{(M+1)}{(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_p^{(m)} - \hat{\beta}_p\right)^2}\right)^2 \tag{3}$$

where $m = 1, \ldots, M$ imputations are taken from draws widely separated to practically eliminate autocorrelation. Multivariate combining rules for the joint distribution of $\hat{\beta}$ are available as well (Schafer 1997, 112–118).

1.1.2. Fixed-Effects Model (FX_APR)

Compared to the predictive model using standard generalized linear regression, we can add dummy variables indicating stratum and cluster memberships to account for stratification and clustering effects. Note that we also need to include the log transformation of sampling weight as a predictor if the missing-data mechanism depends on weights to make the imputation model truly appropriate. The model takes the following form:

$$g\big(E(Y_i|X_i)\big) = X_i\beta + D_i\gamma + E_i\eta + [\zeta\log(w_i)], \tag{4}$$

where D_i is a $1 \times (H-1)$ row vector of dummies representing the H strata, and E_i is a $1 \times Q$ row vector of dummies representing the clusters nested within each stratum. Note that $Q = \sum_h Q_h - H$, where Q_h is the number of clusters in each stratum; in the case of the two-PSU-per-stratum case, $Q = H$. The parameter space under this model is expanded as $\theta = (\beta,\gamma,\eta,\zeta)$, and the steps for imputation are similar to those in the SRS setting.

1.1.3. Mixed-Effects Model (RE_APR)

As there are only two PSUs selected from each stratum, it is not feasible to model clusters as random effects separately within each stratum. Here we pool all $Q + H$ clusters in the sample and model them using a single random-effect term. The imputation model is specified as follows:

$$g\big(E(Y_j|X_j)\big) = X_j\beta + D_j\gamma + u_i + [\zeta \log(w_j)], \tag{5}$$

where $u_i \sim N\big(0, \sigma_u^2\big)$ is a random intercept term representing cluster effects, for $i = 1, \ldots, (Q + H)$, and σ_u^2 denotes the between cluster variance. Other terms are as previously defined. (In the two-PSU-per-stratum case, $Q + H = 2H$.)

2. Synthetic MI Using the Weighted FPBB for Stratified Samples

In this section, we develop the two-step multiple-imputation methodology for a stratified two-stage sample design where a combination of complex sampling techniques are considered, namely, stratification, clustering, and unequal inclusion probability. We develop methods for an unrestricted number of clusters per stratum, but for our simulations and application we focus on the special case of two PSUs selected per stratum, which mimics the form of a public-use dataset that is commonly released for analyses.

2.1. Synthetic Data Generation to Account for Complex Sample Designs

Consider a finite population P, which is stratified into H strata with N_h PSUs in the h^{th} stratum, and hence the population size of PSUs is $\sum_{h=1}^{H} N_h = N$. For the h^{th} stratum, select n_h PSUs with/without replacement from some probability sampling plan, independently across strata, and hence the total sample size of PSUs is $\sum_{h=1}^{H} n_h = n$. Subsampling of m_{hi} elements (treated as the ultimate sampling units in this example) from a total of M_{hi} is then conducted within the i^{th} sampled PSU of the h^{th} stratum for $i = 1, \ldots, n_h, h = 1,2, \ldots, H$. Hence the overall sample size and population size of elements are $\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi} = \sum_{h=1}^{H} m_h = m$ and $\sum_{h=1}^{H}\sum_{i=1}^{N_h} M_{hi} = \sum_{h=1}^{H} M_h = M$, respectively, where m_h and M_h are the sample size and population size of elements for the h^{th} stratum, respectively. The population consists of four types of survey variables: a single outcome Y, a single covariate X, a design matrix $Z = [S,C,w]$ including the stratum indicators (S), the cluster indicator (C) and the sample weight (w), and the response indicator R. Let $D = (D_s, D_{ns}) = \{(Y_{hij}, X_{hij}, Z_{hij}, R_{hij}), h = 1, \ldots, H, i = 1, \ldots, N_h, j = 1, \ldots, M_{hi}\}$ denote the population of values measured on the survey variables, which is divided into the sampled component (D_s) and the nonsampled (D_{ns}) component.

We generate synthetic populations using a two-stage procedure. The first stage accommodates stratification and clustering and the second weighting. We have two broad approaches. The first, which we term SYN1, assumes that first-stage (cluster-level) and second-stage (element-level) sample weights are available for the analysis and implements a weighted FPBB at each level to generate the synthetic population. The second, which we term SYN2, assumes that only final weights are available for the analysis; it uses a Bayesian bootstrap to account for stratification and clustering at the first stage and the weighted FPBB to account for the final weight at the second stage.

2.1.1. Double-Weighted Finite Population Bayesian Bootstrap (SYN1)

For the h^{th} stratum, let $t_{s,h}$ and $t_{ns,h}$ index the sampled and nonsampled clusters, respectively, and $\{b^1, \ldots, b^q, \ldots, b^{r_h}, q = 1, \ldots, r_h\}$ be the r_h $(1 \leq r_h \leq N_h)$ distinct matrices of real numbers each of dimension $|b^q_{row}| \times |b^q_{col}|$ with no row vectors in common. Each cluster in the stratum can take the form of one of b^qs. Let $t_{hi} = q$ when the i^{th} cluster takes on the values of b^q, for $i = 1, \ldots, N_h$. Assume $n_h = r_h$ and $m_{hi} = \|b^{t_{s,hi}}\|$ (the number of distinct row vectors in $b^{t_{s,hi}}$) for convenience of exposition. Let $w_{t_{s,h}}(i)$ be the sample weight of the i^{th} sampled cluster in the h^{th} stratum which equals b^q, for $i = 1, \ldots, n_h$. Also let $w_{t_{s,hi}}, D_{s,h}(j)$ be the sample weight of the j^{th} sampled element in the i^{th} sampled cluster which equals $b_k^{t_{s,hi}}$, for $j = 1, \ldots, m_{hi}$. Finally, let $c_{t_{s,h}}(q)$ and $c_{t_{ns,h}}(q)$ be the number of sampled and nonsampled clusters that equal b^q, and $c^{hi}_{t_h,D_{s,h}}(k)$ and $c^{hi}_{t_h,D_{ns,h}}(k)$ be the number of sampled and nonsampled elements that equal $b_k^{t_{s,hi}}$.

It can be shown (cf. Zhou 2014) that, *within a stratum h*, the Polya posterior for the counts of distinct unobserved elements $D_{ns,h}$ is given by

$$p\left(D_{ns,h}|D_{s,h}\right) = \frac{\left\{\prod_{q=1}^{r_h}\left\{\Gamma(w_{t_h'}(q))/\Gamma(w_{t_{s,h}}(q))\right\}\right\}}{\left\{\Gamma(N_h)/\Gamma(n_h)\right\}} \times \frac{\left\{\prod_{k=1}^{m_h}\left\{\Gamma(w_{t_h',D_{ns,h}}(k))/\Gamma(w_{t_{s,h},D_{s,h}}(k))\right\}\right\}}{\left\{\Gamma(M_h)/\Gamma(m_h)\right\}}, \tag{6}$$

where $w_{t_h'}(q) = w_{t_{s,h}}(q) + c_{t_{ns,h}}(q)$ and $w_{t_h',D_{ns,h}}(k) = w_{t_{s,h},D_{s,h}}(k) + c^{hi}_{t_h,D_{ns,h}}(k)$, for $m_h = \sum_{k=1}^{m_h} c^{hi}_{t_h,D_{s,h}}(k)$ and $m_h' = M_h - m_h = \sum_{k=1}^{m_h} c^{hi}_{t_h,D_{ns,h}}(k)$. The *full posterior* is then given by the product of the posteriors within each stratum, since these strata are independent and all strata in the population are in the sample:

$$p\left(D_{ns}|D_s\right) = \prod_{h=1}^{H} p\left(D_{ns,h}|D_{s,h}\right). \tag{7}$$

A Monte Carlo procedure to simulate from this posterior distribution is then given as follows:

(i) Draw the $N_h - n_h$ nonsampled clusters in the population based on the Polya posterior distribution independently for each stratum. Each of the sampled clusters is resampled with probability

$$\varsigma_{hi} = \frac{w_{t_{s,h}}(i) - 1 + l_{hi,k-1} \times \left(\dfrac{N_h - n_h}{n_h}\right)}{N_h - n_h + (k-1) \times \left(\dfrac{N_h - n_h}{n_h}\right)}, k = 1, \ldots, N_h - n_h + 1, \tag{8}$$

where $l_{hi,k-1}$ is the number of times that the i^{th} cluster in the h^{th} stratum has been resampled at the $(k-1)^{th}$ resampling, and $w_{t_{s,h}}(i)$ is the weight for the i^{th} sampled cluster in the h^{th} stratum which is normalized to sum up to the total number of clusters, that is, $\sum_{i=1}^{n_h} w_{t_{s,h}}(i) = N_h$.

(ii) From Step 1, form a population of clusters $\{c_{11}, c_{12}, \ldots, c_{1n_1}, \overset{*}{c}_{11}, \overset{*}{c}_{12}, \ldots, \overset{*}{c}_{1N_1-n_1}, \ldots, c_{H1}, c_{H2}, \ldots, c_{Hn_H}, \overset{*}{c}_{H1}, \overset{*}{c}_{H2}, \ldots, \overset{*}{c}_{HN_H-n_H}\}$. Record the number of times each of the clusters from the original sample appears in the FPBB population of clusters, denoted by $\tau_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, H.$, and $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\tau_{hi} = N$. Then update the within cluster *element-level conditional weights* as follows: $\overset{*}{w}_{j|hi} = w_{j|hi} \times \tau_{hi}$,

$i = 1, \ldots, n_h, h = 1, \ldots, H$, where $w_{j|hi}$ is the inverse of the conditional probability that element j is selected given cluster i in stratum h is selected. Now pool all elements from these clusters together and treat them as a single *FPBB sample* (i.e., as if they have no stratum or cluster boundaries). Note that this FPBB sample has the same sample size $m = \sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$ as the original sample, but different sampling weights. We then once more apply the weighted FPBB to these pooled elements to generate $M - m$ units from the m units in the FPBB sample. We resample from each of the resampled clusters $M - m$ elements, cycling through $M - m$ times and resampling with probability

$$\lambda_{j|hi} = \frac{w^{*}_{j|hi} - 1 + l_{hij,k-1} \times \left(\frac{M-m}{m}\right)}{M - m + (k-1) \times \left(\frac{M-m}{m}\right)}, \quad k = 1, \ldots, (M - m + 1), \qquad (9)$$

where $l_{hij,k}$ is the number of times that the j^{th} element in the i^{th} cluster in the h^{th} stratum has been resampled at the k^{th} resampling, and $w_{j|hi}$ is the updated conditional weight for the j^{th} element in the i^{th} cluster in the h^{th} stratum. Again, they are normalized to sum up to the total number of units in the entire population, that is, $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{j|hi} = M$. Thus we create a single synthetic population. Repeat Step 2 B times to obtain B FPBB synthetic populations.

(iii) Repeat Steps 1-2 L times to obtain L bootstrap samples, yielding $L \times B$ FPBB populations $P^{syn}_{(lb)} = \left(P^{syn}_{(lb)obs}, P^{syn}_{(lb)mis}\right)$, $l = 1, \ldots, L$, $b = 1, \ldots B$, each of which consists of both responding elements and nonresponding elements on a vector of variables $\{Y, X, Z, R\}$.

2.1.2. Bootstrap — Weighted Finite Population Bayesian Bootstrap (SYN2)

Because we often do not know the first- and second-stage weights in public-use datasets, we consider an alternative to the procedure proposed in Subsection 2.1.1. Rather than obtaining a sample of clusters from a draw from a Polya posterior, we use replication methods (Rust and Rao 1996) to capture the cluster-level sampling variance. The final sampling weights instead of the adjusted element-level conditional weights are then used directly as input in the second-stage weighted FPBB. We use Rao and Wu's (1988) rescaling bootstrap, which is a generalized extension of McCarthy and Snowden's (1985) "with replacement bootstrap". Once the PSUs have been sampled, we continue with the weighted FPBB approach to complete the synthetic population data generation. The proposed procedure is as follows:

(i) Select a sample of $n^{*}_h = n_h - 1$ PSUs from the parent sample in each stratum via SRSWR sampling;

(ii) Apply the "ultimate cluster principle" (Wolter 2007), that is, once a PSU is taken into the bootstrap replicate, all elements in that PSU are taken into the replicate also. Thus we obtain our first bootstrap sample;

(iii) Repeat the previous steps L times to obtain L bootstrap samples $\{Boot\_l, l = 1, \ldots, L\}$;

(iv) Within each bootstrap sample, update the element-level sampling weights as:

$$w^*_{hij} = w_{hij} \times \left(\tau_{hi}\frac{n_h}{n^*_h}\right) = \begin{cases} = \frac{n_h}{n_h-1} w_{hij}, \text{ if the } i^{th} \text{ PSU selected in the bootstrap sample} \\ = 0, \text{ if the } i^{th} \text{ PSU not selected in the bootstrap sample} \end{cases}$$

As w^*_{hij} itself implicitly carries over the strata and PSU information in addition to unequal inclusion probability, we can drop the subscripts *hi* henceforth by pooling all elements in the bootstrap sample regardless of which stratum and PSU they originally came from. Normalize w^*_js to sum up to m^*: $\sum_{j=1}^{m^*} w^*_j = m^*$, where m^* is the bootstrap sample size.

(v) For the l^{th} bootstrap sample, $l = 1, \ldots, L$, apply the weighted FPBB algorithm to create an entire population $D = \left(D_{ns}, D^*_s\right)$ based on the posterior predictive distribution of elements in the nonsampled population $D_{ns} = \left\{(Y_j, X_j, Z_j, R_j), j = m^* + 1, \ldots, M\right\}$ given the elements in the bootstrap sample $D^*_s = \left\{(Y_j, X_j, Z_j, R_j), j = 1, \ldots, m^*\right\}$.

Operationally, we draw a Polya sample of size $M^* = M - m^*$ from $mult\left(M^*; \lambda_1, \ldots, \lambda_K\right)$ where the selection probability $\lambda_k, k = 1, \ldots, K$ is a function of w^*_j:

$$\lambda_k = \frac{w^*_j - 1 + l_{j,k-1} \times \left(\frac{M^*}{m^*}\right)}{M^* + (k-1) \times \left(\frac{M^*}{m^*}\right)}, k = 1, \ldots, M^* + 1, \tag{10}$$

Repeat Step (v) for B times to obtain $L \times B$ FPBB populations.

2.2. *Imputation of the Synthesized Populations*

Once the set of FPBB synthetic populations $P^{syn} = \left\{P^{(l)}_{(b)}, l = 1, \ldots, L, b = 1, \ldots, B\right\}$, where $P^{(l)}_{(b)} = \left(Y^{(l)}_{(b)mis}, P^{(l)}_{(b)obs}\right)$ are created using either the SYN1 method or the SYN2 method, we generate imputations $P^{imp} = \left\{P^{(l)}_{(ba)}, l = 1, \ldots, L, b = 1, \ldots, B, a = 1, \ldots, A\right\}$ from the posterior predictive distribution $p\left(Y^{(l)}_{(b)mis} | P^{(l)}_{(b)obs}\right)$ based on a parametric model that does not condition on sample design features, that is, a model taking a form similar to the SRS model given in Subsection 2.1. We consider imputations based on the covariate (X) only (SYN1_srs or SYN2_srs) or imputations that include the log of the sample weights in the linear predictors (SYN1_lwt or SYN2_lwt).

To obtain the MI inference, denote the observed set of synthetic populations by $P_R = \left\{P^{(l)}_{(b)obs}, b = 1, \ldots, B, l = 1, \ldots, L\right\}$ and the imputed set of synthetic populations by $P_{\bar{R}} = \left\{Y^{(l)}_{(ba)mis}, l = 1, \ldots, L, b = 1, \ldots, B, a = 1, \ldots, A\right\}$. The MI point estimator for the population statistic of interest Q (mean, regression estimator, quantile) is then given by the mean of the lba^{th} point estimators:

$$\hat{Q}_{MI} = \frac{1}{LBA} \sum_l \sum_b \sum_b \hat{Q}_{lba}. \tag{11}$$

The MI variance estimator is:

$$\hat{V}_{MI} = (1 + L^{-1})V_L = (1 + L^{-1})\frac{1}{L-1}\sum_l \left(\hat{Q}_l - \hat{Q}_{MI}\right)^2, \text{ where} \tag{12}$$

$$\hat{Q}_l = \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}.$$

We then construct the 95% interval estimate for quantiles based on t reference distribution with degrees of freedom equal to $\min\left\{v_{com} = \sum_h n_h - H, v_{syn} = L - 1\right\}$. These results arise from the fact that, by the standard Rubin (1987) MI combining rules, we have

$$Q|P^{imp} \dot{\sim} t_{L-1}\left(\bar{Q}_L, (1 + L^{-1})V_L\right), \tag{13}$$

where $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$, $V_L = \frac{1}{L-1}\sum_l \left(\tilde{Q}^{(l)} - \bar{Q}_L\right)^2$, and $\tilde{Q}^{(l)} = \lim_{\substack{B\to\infty \\ A\to\infty}} \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}$.

Replacing $\tilde{Q}^{(l)}$ with its finite simulation estimator $\hat{Q}_l$ replaces $\bar{Q}_L$ with $\hat{Q}_{MI}$ and gives the results above. A complete theoretical justification for (13) is provided in Dong et al. (2014) and Zhou (2014). Some intuition of the result can be gained by noting that the generation of the synthetic population sets the within imputation variance to 0 so that the posterior variance of Q can be obtained using the between-bootstrap variance only. Moreover, (11) assumes that $E(\hat{q}_{ba}) = Q$ – a result guaranteed by our Bayesian bootstrap estimator if the imputation model is also correct – as well as a sufficiently large sample size for the t approximation is reasonable.

Lo (1988) showed that the variance estimator for the FPBB mean in a simple random sample setting should be inflated by the factor $(\frac{n+1}{n-1})$. In the double-weighted FPBB (SYN1) setting, a small sample correction to the variance estimate thus needs to be used when the number of clusters per stratum is small. When $n_h = a$ is a constant across all strata, we use $\frac{n_h+1}{n_h-1}(1 + L^{-1})\ V_L$; otherwise we suggest $\frac{\bar{n}_h+1}{\bar{n}_h-1}(1 + L^{-1})\ V_L$, where $\bar{n}_h = H^{-1}\sum_h n_h$.

The Appendix provides the sample R code used to conduct the analyses in the application in Section 4 and can easily be adapted to other settings.

3. Simulation Study

We conducted a simulation study to investigate the performance of the proposed method for incorporating stratified cluster-sampling effects in multiple imputation. We targeted three population statistics: 1) population quantiles, 2) proportions of binary event data, and 3) logistic regression parameters relating the covariate to the binary data. The simulation is a 2 × 2 factorial design based on the following factors:

1) keeping the first-stage sampling plan constant, we let the subsampling rate f_2 of elements within sampled clusters be
   a) independent of or
   b) dependent on the stratum effects, and
2) assume that
   a) the missingness on the Y-variable (continuous or binary) depends only on the covariate (X) (MAR_X), or
   b) depends on both X and the final sampling weight W(MAR_X,W).

We focus on a two-PSU-per-stratum sample design, both because it is a common design, especially in public-use settings, and because it is a “limiting case” in terms of the number of PSUs per stratum. In addition to the two variants of our synthetic MI estimators, we consider standard parametric MI under the SRS, appropriate fixed-effect (FX_APR), and appropriate random-effect (RE_APR) models.

3.1. Data Generation

Let i be the index for strata, j be the index for clusters, and k be the index for elements. Suppose there are 50 strata in the population. First, the number of PSUs in each stratum is randomly determined according to a uniform distribution, that is, $C_i \sim Unif(2{,}54)$, $i = 1,\ldots,50$; second, the number of population elements within PSUs is randomly generated as $N_{ij} \sim Unif(20{,}80)$, $i = 1,\ldots,50, j = 1,\ldots,C_i$. Thus we obtain a population of size $N = 67385$. The complete data for four survey variables $Y = (Y_1,Y_2,Y_3,Y_4)^T$ are generated from a superpopulation model according to a two-step process, In the first step, Y_1 and Y_2 are randomly selected from a bivariate linear mixed-effects model; let $N_2(\cdot)$ denote a bivariate normal distribution function:

$$\begin{pmatrix} Y_{1ijk} \\ Y_{2ijk} \end{pmatrix} \sim N_2(\mu,\Sigma), \text{ where } \mu = \begin{bmatrix} \beta_1 + S_i + u_{1ij} + \varepsilon_{1ijk} \\ \beta_2 + u_{2ij} + \varepsilon_{2ijk} \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}. \quad (14)$$

Let $\beta_1 = \beta_2 = 15$ be the fixed covariate effects, $S_i = \frac{i}{5}$ be the fixed stratum effects, and let $\begin{bmatrix} u_{1ij} & u_{2ij} \end{bmatrix}^T$ and $\begin{bmatrix} \varepsilon_{1ijk} & \varepsilon_{2ijk} \end{bmatrix}^T$be the random cluster effects and random error terms drawn from two independent bivariate normal distributions: $N_2(0,\Sigma_u)$ and $N_2(0,\Sigma_\varepsilon)$. Elements of Σ_u are set as: $\sigma^2_{u_1} = 4$, $\sigma^2_{u_2} = 1$, $\sigma_{u_1u_2} = 0.2$, and elements of Σ_ε are set as: $\sigma^2_{\varepsilon_1} = 4$, $\sigma^2_{\varepsilon_2} = 3$, $\sigma_{\varepsilon_1\varepsilon_2} = 1.732$. This results in conditional intraclass correlations (ICC) of Y_1 and Y_2 as $\rho_{Y_1} = 0.5$ and $\rho_{Y_2} = 0.25$ (note that the unconditional ICC for the two variables may be smaller than these values). In the second step, a random-effects logistic regression model (Anderson and Aitkin 1985; Stiratelli, et al. 1984) is used to simulate two binary outcome variables Y_3 and Y_4 as a function of Y_2. Under this model, a random effect is added to the linear part of the logistic regression model for each element in the cluster. The conditional mean of Y_{3ijk} and Y_{4ijk} is

$$\pi_{ijk} = E\left(Y_{\cdot ijk}|Y_{2ijk}, u_{\cdot ij}\right) = \Pr\left(Y_{\cdot ijk} = 1|Y_{2ijk}, u_{\cdot ij}\right) = \frac{e^{\alpha_0+\alpha_1 S_i+\alpha_2 Y_{2ijk}+u_{\cdot ij}}}{1+e^{\alpha_0+\alpha_1 S_i+\alpha_2 Y_{2ijk}+u_{\cdot ij}}}, \quad (15)$$

where $u_{3ij} \sim N(0{,}6^2)$, $u_{4ij} \sim N(0{,}10^2)$ and $\alpha = (\alpha_0,\alpha_1,\alpha_2)^T$ is the vector of fixed covariate effects. We fix $\alpha_2 = 1.5$ and vary α_0 and α_1 to obtain two different binary variables Y_{3ijk} and Y_{4ijk}, with either moderate $(\alpha_0 = -5,\alpha_1 = -1.5)$ or rare probabilities $(\alpha_0 = -8,\alpha_1 = -6)$. Given $u_{\cdot ij}$, the $Y_{\cdot ijk}s$ in the cluster are independent Bernoulli variables, that is, $Y_{\cdot ijk}|u_{\cdot ij} \sim Bern(\pi_{ijk})$.

Figure 1 shows the correlations between variables in the simulated population, with the different shades of grey representing different degrees of association between any of the two variables. The darker shades indicate higher correlation. All survey outcome variables (Y_1,Y_3,Y_4) have a moderate to strong (0.2∼0.8) stratum effect (H or *strID*) and clustering effect (U_1,U_3,U_4), indicating that accounting for these effects in the analysis of missing data is essential.

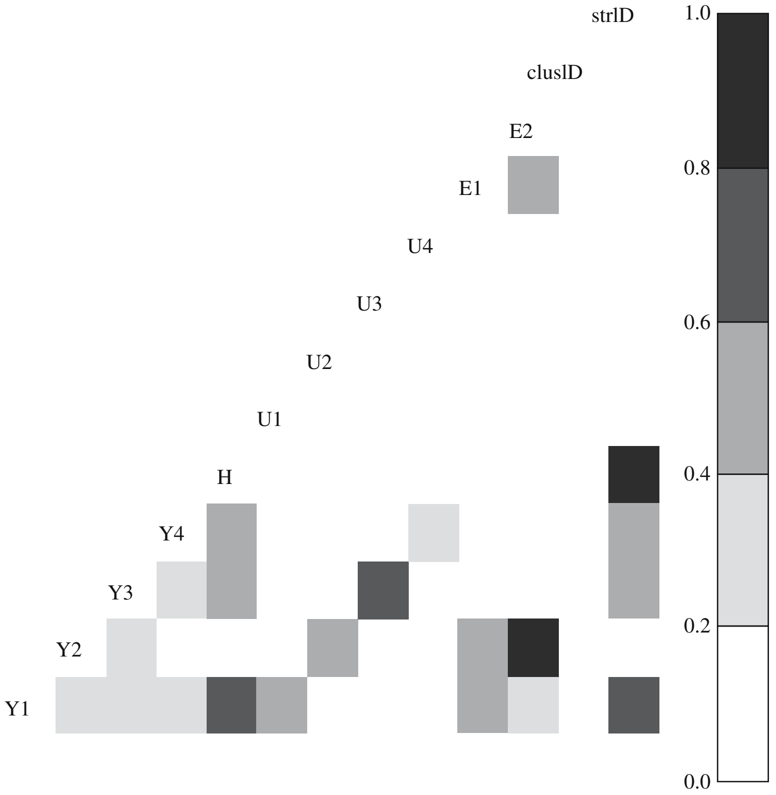


Fig. 1. Correlation between variables in the simulated population (darker shades = higher correlation)

3.2. Sample Design

Within each stratum, we draw a two-stage cluster sample according to the following procedure: first, we draw a sample of two PSUs without replacement with probability proportional to the cluster size $f_{1ij} = \frac{2^{*}N_{ij}}{\sum_{j} N_{ij}}$. Second, we sample elements from each sampled cluster using two different subsampling schemes:

1) sampling probability independent of S_i which is defined in (14): SRS with an equal sampling fraction of $f_{2k|ij} = 1/5$; and
2) sampling probability related to S_i: SRS with varying sampling fractions across strata, that is $f_{2k|ij} = \text{expit}(-0.8 - 0.12^{*}S_i)$, where $\text{expit}(x) = 1/(1 + \text{e}^{-1}(x))$.

An average of 1,122 elements are selected in each of the 200 simulation replications. The distributions of sampling weights are shown in Figure 2. The distributions of sampling weights under the two subsampling schemes are generally very similar with somewhat more skewness under subsampling scheme 2.

3.3. Imposing Missingness

Throughout the simulation study, we assume that Y_2 is always completely observed and we impose missing values on Y_1, Y_3, and Y_4 independently according to the following deletion

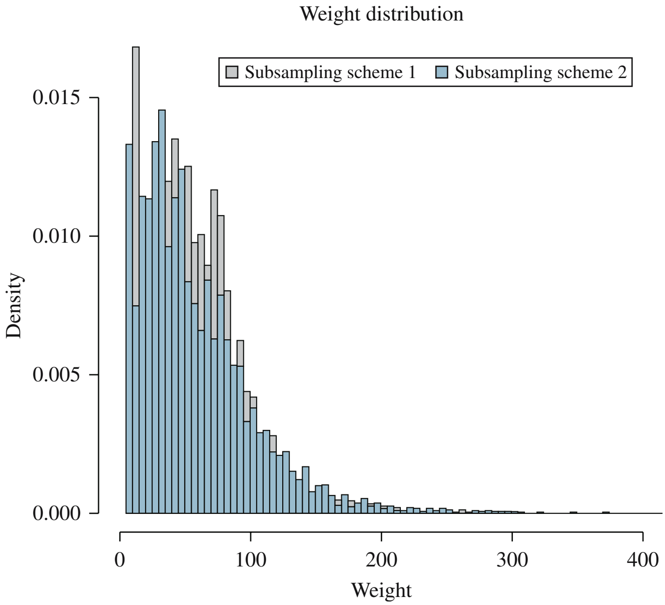


Fig. 2. Distribution of weights under the two subsampling schemes

function conditional on Y_2 and/or log transformation of the weight:

$$\Pr\left(R=0|Y_2,W\right)=\frac{\exp\left(\lambda_0+\lambda_1{*}Y_2+\lambda_2{*}\log\left(W\right)\right)}{1+\exp\left(\lambda_0+\lambda_1{*}Y_2+\lambda_2{*}\log\left(W\right)\right)},\tag{16}$$

where R is the response indicator and W is the overall sample weight. Setting $\lambda_2=0$, we obtain the first MAR mechanism (i.e., MAR_X, note that we treat Y_2 as a covariate X here), under which we further set $\lambda_0=3.42$, $\lambda_1=-0.2$ and $\lambda_0=-2.58$, $\lambda_1=0.2$ for deleting values on Y_1 and Y_3, Y_4, respectively. Setting $\lambda_2=-0.6$, we obtain the second MAR mechanism (i.e., MAR_X,W), under which we fix $\lambda_1=0.2$ and set two values on $\lambda_0(=-0.274$ or $-0.33)$ for deleting values independently on all three outcome variables under subsampling scheme 1 and subsampling scheme 2, respectively. All deletion functions result in approximately 40% missingness on each variable.

3.4. *Parametric Multiple Imputation*

Both simple random sample SRS (including SRS, SYN1_srs and SYN2_srs) and fixed-effects model FX_APR can be implemented in R (R Core Team 2013) using *mice* routines; for the logistic model associated with the binary outcome, the method '*logreg*' must be specified. We use the *pan* package in R for the mixed-effects imputation (RE_APR) for the missing continuous outcome; logistic mixed-effects imputation is programmed in SAS for the missing binary outcome, as there is no missing-data software package readily available for use.

3.5. Parameters of Interest and Inference

We focus on inference for the following population parameters: the mean of the continuous variable Y_1, the mean of the binary variables Y_3 and Y_4 (i.e., Bernoulli proportions), linear regression coefficients of Y_1 on Y_2, logistic regression coefficients of Y_3 (or Y_4) on Y_2, and the population percentiles of the continuous variable Y_1.

Weighted analyses and sandwich variance estimators accounting for strata and clusters are used to estimate smooth statistics (including proportions and regression parameters) under the three fully parametric MI methods. For estimating quantiles of the distribution of a continuous survey variable, we construct the sample-weighted point estimator with confidence intervals based on the test-inversion method (Francisco and Fuller 1991). We chose the test-inversion method instead of Woodruff's method (Woodruff 1952) despite the computational intensity, because the literature suggests that it may outperform Woodruff in heavily stratified samples or in small-to-moderate-sized samples (Kovar et al. 1988). Based on the a^{th} imputed dataset, the empirical distribution function can be written as

$$\hat{F}^{(a)}(y) = \frac{\left[\sum_{S_R} w_{hij} I\left(y_{hij}^{obs} < y\right) + \sum_{S_{\bar{R}}} w_{hij} I\left(y_{hij}^{(a)} < y\right)\right]}{\sum_{S} w_{hij}}, \qquad (17)$$

where S_R and $S_{\bar{R}}$ are subsets of the sample data S, consisting of respondents and nonrespondents respectively. The estimator $\hat{F}(y)$ and its associated estimated variance $v(\hat{F}(y))$ can then be obtained using the variance estimator proposed by Francisco and Fuller (1991) together with standard Rubin combining rules as previously described. The sample γ^{th} quantile estimator thus is $\hat{q}_\gamma = (\hat{F})^{-1}(\gamma)$, with 95% asymptotic confidence interval (CI) given by

$$[L, U] = \left[[\hat{F}]^{-1}\left(\gamma - t_{0.025}\sqrt{\operatorname{var}(\hat{F}(q_\gamma))}\right), [\hat{F}]^{-1}\left(\gamma + t_{0.025}\sqrt{\operatorname{var}(\hat{F}(q_\gamma))}\right)\right]. \qquad (18)$$

3.6. Results

Table 1 compares the average width × 10^{-2} and average coverage rates of the 95% CI of $q(\alpha)$, where $\alpha = 0.05$, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.95, corresponding to seven selected population quantiles. Among all methods considered, the SRS imputation model yields the poorest coverage. This results from the compounding effects of biases and variance underestimation, due to ignoring stratum effects and clustering effects respectively. As we increase the dependence of both the sampling mechanism and response mechanism on stratum effects and sampling weights, the performance of SRS becomes even worse, as exhibited by the markedly increased RelBias and decreased coverage rates. In addition, ignoring stratum and/or weight effects that are highly relevant to either mechanism seems to impact the median and second and third quartiles more than the tail quantiles under SRS, as evident in the relatively lower coverage rates in the right part of Table 1.

Table 1. Comparison of average width $\times 10^{-2}$ *and 95% CI coverage rates of* $q(\alpha)$ *for* $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90,$ *and 0.95.*

Sampling scheme	Missingness mechanism	Methods	Average width of 95% CI x 10^{-2}							95% CI coverage						
			0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$f_2 \propto$ const.	Complete data	Actual	170	144	123	106	116	142	165	90.5%	92.5%	94.5%	93.5%	95.5%	91.5%	91.0%
		Syn1BD	172	144	126	105	117	146	165	90.5%	90.0%	95.0%	94.0%	95.0%	94.0%	89.0%
		Syn2BD	182	154	132	113	122	150	171	94.0%	95.0%	96.0%	94.5%	96.5%	96.0%	92.5%
	MAR_X	SRS	165	134	112	101	108	132	158	93.0%	91.5%	86.0%	82.5%	83.0%	89.0%	93.5%
		FX_APR	171	143	120	105	116	146	172	92.5%	90.5%	90.5%	92.5%	93.5%	94.0%	95.0%
		RE_APR	184	154	131	115	125	156	186	93.5%	94.0%	93.0%	97.5%	95.5%	95.5%	97.0%
		Syn1_srs	171	145	123	109	122	148	165	91.0%	89.5%	92.5%	95.0%	90.5%	89.5%	94.0%
		Syn2_srs	182	158	134	118	129	156	175	93.5%	93.0%	94.5%	96.5%	94.5%	94.5%	95.0%
	MAR_X,W	SRS	178	146	120	109	110	139	163	89.0%	81.0%	70.5%	69.0%	80.0%	90.0%	91.0%
		FX_APR	186	153	126	115	125	155	190	89.5%	92.5%	93.5%	95.5%	92.5%	92.5%	96.0%
		RE_APR	197	166	140	127	136	168	197	95.0%	97.0%	97.0%	98.0%	96.0%	95.0%	96.0%
		Syn1_srs	173	150	124	111	119	146	163	91.5%	92.0%	93.0%	91.5%	90.0%	94.0%	92.5%
		Syn2_srs	183	160	134	119	123	153	172	93.5%	95.5%	96.5%	92.5%	92.0%	93.0%	95.5%
		Syn1_lwt	174	151	126	115	124	148	166	90.0%	89.0%	93.0%	94.5%	90.5%	96.0%	94.0%
		Syn2_lwt	184	161	136	122	132	155	174	92.0%	93.0%	95.5%	96.0%	94.5%	96.0%	95.0%
$f_2 \propto h(S_i)$	Complete data	Actual	170	143	120	110	121	148	169	92.5%	94.5%	95.0%	96.0%	92.5%	87.5%	87.5%
		Syn1BD	177	142	120	108	121	152	175	91.0%	92.5%	92.0%	94.5%	92.5%	87.5%	87.5%
		Syn2BD	182	152	128	116	126	154	178	95.0%	97.0%	96.0%	97.0%	94.5%	90.0%	90.5%
	MAR_X	SRS	175	139	121	111	116	141	169	86.5%	73.0%	57.0%	48.5%	61.0%	72.0%	80.5%
		FX_APR	174	142	121	113	124	162	202	95.5%	95.0%	98.0%	95.5%	93.5%	92.5%	95.5%
		RE_APR	181	150	128	119	131	168	205	94.0%	96.5%	97.0%	96.5%	97.0%	94.0%	96.0%
		Syn1_srs	166	140	119	111	126	156	180	93.5%	94.0%	96.5%	92.5%	92.0%	91.0%	90.0%
		Syn2_srs	179	152	129	119	132	162	185	94.5%	95.5%	98.0%	96.5%	95.0%	93.5%	92.5%
	MAR_X,W	SRS	191	157	127	117	122	147	168	47.0%	31.5%	9.5%	8.0%	30.0%	60.0%	73.5%
		FX_APR	186	153	125	119	138	179	227	96.5%	97.0%	93.5%	96.5%	97.0%	95.0%	94.5%
		RE_APR	190	161	135	131	148	184	220	98.0%	99.5%	97.5%	98.0%	98.5%	96.5%	95.0%
		Syn1_srs	168	146	124	114	128	155	174	94.0%	92.5%	84.0%	73.0%	76.0%	87.5%	88.0%
		Syn2_srs	184	160	135	122	134	160	179	95.0%	95.5%	88.0%	77.5%	79.0%	87.0%	89.0%
		Syn1_lwt	168	143	121	113	130	158	176	92.5%	92.5%	94.5%	92.0%	89.0%	92.0%	91.5%
		Syn2_lwt	178	155	131	122	138	166	185	96.0%	95.5%	95.5%	92.5%	95.5%	95.0%	93.0%

The FX_APR model (Reiter et al. 2006; Rubin 1996; Schenker et al. 2006), generally performs fairly well in our simulation study with respect to the estimation of population quantiles. There is some modest underestimation of the small percentile quartiles with the second-stage sampling constant. The RE_APR model also performs well, with the exception of moderate to high overcoverage when the second-stage sampling probability is associated with the stratum mean and the missingness mechanism.

In contrast, our synthetic MI (SYN2 in particular) compares favorably with all of its competitors, and in most cases yields results comparable to the RE_APR, which is regarded as a "gold standard" as it is compatible with the data-generating mechanism (Meng 1994). There is some undercoverage when the stratified double-weighted FPBB estimator (SYN1) is used, perhaps due to the fact that the Lo small-sample adjustment is not as accurate when $n_h = 2$. However, use of a stratified bootstrap-weighted FPBB estimator (SYN2) generally eliminates this issue. Although an imputation model assuming SRS suffices for the synthetic MI method in most scenarios, we need to include the sampling weight as a predictor when the outcome Y and the response indicator R are strongly associated with each other through the sampling mechanism I, as is the case with the second subsampling scheme, when both the missingness indicator and the second-stage sampling rate are functions of the stratum mean.

Tables 2 and 3 compare the absolute relative bias $relbias = 100 \times \frac{|\hat{\theta} - \theta_{complete}|}{\theta_{complete}}\%$, RMSE and 95% nominal CI coverage for the estimated mean/proportions of Y_1, Y_3 and Y_4 and the slopes of the three outcome variables on Y_2, respectively. ($\theta_{complete}$ is the estimated parameter with complete data, and $\hat{\theta}$ is the estimated parameter under one of the different MI methods.) As in the estimation of the quantiles, the SRS imputation model is biased and has poor coverage as it ignores stratum and cluster effects. Again, dependence of subsampling on stratum effects and dependence of response on sampling weights damage the performance of SRS even further.

FX_APR generally performs well in estimating the mean of a continuous variable (Y_1) and a regular binary variable (Y_3) with moderate probability as well as the slopes. However, it fails for proportion estimation for rare events data (Y_4), yielding biased point estimates and less than nominal coverage throughout all scenarios. One interpretation might be that overfitting occurs when too many dummies are included to account for fixed strata and cluster effects, yielding dummy variables where all observed cases are 0 or 1. In this case, "complete separation" yields unstable coefficient estimates, damaging the predictive efficacy when the fitted model is used for drawing missing values. The problem is particularly prominent when the logistic fixed-effects imputation model is used along with the current sampling design, where an average of only ten elements are selected per PSU within each stratum; this results in even more substantial biases on $\bar{Y}_4$ than the SRS model. (Use of a Bayesian approach with an informative prior of the form $t_1(0,2.5)$ on the fixed-effect parameters using the *mi* function in R (Gelman et al. 2008) reduced but did not remove the impact of complete separation. A relative bias of $12-13\%$ remained for the estimation of of $\bar{Y}_4$ under the MAR_X missingness mechanism, with 95% nominal coverage of 89%, while a relative bias of $17-22\%$ remained under the MAR_X,W mechanism, with nominal coverage of 84%.) The random-effects model RE_APR more effectively avoids the overfitting issue through shrinkage effects: note that under RE_APR, we pooled all PSUs from all

Table 2. Comparison of RelBias, RMSE and 95% CI coverage rates for the mean of Y1 and proportions of Y3 and Y4, Population true value: $\bar{Y}_1 = 20.4$, $P_{Y_3} = 0.608$, $P_{Y_4} = 0.117$

			RelBias			RMSE			95% CI coverage		
Sampling scheme	Missingness mechanism	Methods	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}
	Complete data	Actual	–	–	–	0.220	0.042	0.024	95.0%	94.0%	90.5%
		Syn1BD	0.0%	0.0%	0.0%	0.221	0.042	0.024	94.5%	94.0%	91.5%
		Syn2BD	0.0%	0.0%	0.0%	0.222	0.043	0.024	95.0%	94.5%	93.0%
$f_2 \propto$ const.		SRS	0.8%	1.6%	10.8%	0.309	0.041	0.028	76.9%	90.0%	85.0%
		FX_APR	0.0%	1.3%	39.2%	0.243	0.040	0.054	91.0%	96.5%	72.5%
Actual samples BD:		RE_APR	0.0%	1.3%	15.1%	0.236	0.040	0.026	93.0%	93.5%	91.0%
$\bar{Y}_1 = 20.3$	MAR_X	Syn1_srs	0.0%	0.3%	0.4%	0.255	0.044	0.025	94.5%	93.5%	91.5%
$P_{Y_3} = 0.604$		Syn2_srs	0.0%	0.2%	0.4%	0.254	0.044	0.025	97.0%	95.0%	94.5%
$P_{Y_4} = 0.117$		SRS	1.4%	2.8%	19.4%	0.398	0.042	0.035	72.0%	85.5%	77.5%
		FX_APR	0.0%	2.7%	48.4%	0.260	0.042	0.065	91.5%	96.0%	60.0%
		RE_APR	0.1%	0.3%	6.8%	0.250	0.041	0.022	97.5%	95.5%	86.0%
		Syn1_srs	0.4%	1.4%	4.2%	0.285	0.043	0.026	92.0%	95.5%	91.5%
	MAR_X,W	Syn2_srs	0.5%	1.4%	4.4%	0.283	0.043	0.026	96.5%	95.0%	96.0%
		Syn1_lwt	0.0%	0.6%	0.3%	0.273	0.045	0.027	95.5%	93.5%	89.0%
		Syn2_lwt	0.0%	0.5%	0.0%	0.271	0.045	0.026	96.0%	96.0%	94.0%
	Complete data	Actual	–	–	–	0.218	0.037	0.023	96.0%	97.5%	92.0%
		Syn1BD	0.0%	0.0%	0.0%	0.220	0.037	0.023	93.5%	94.0%	92.0%
		Syn2BD	0.0%	0.0%	0.3%	0.219	0.038	0.023	96.0%	97.0%	94.0%
$f_2 \propto h(S_i)$		SRS	2.4%	4.7%	29.6%	0.540	0.048	0.045	42.0%	80.5%	62.5%
		FX_APR	0.0%	1.5%	42.0%	0.237	0.036	0.058	94.0%	97.0%	70.5%
Actual samples BD:		RE_APR	0.2%	1.6%	16.1%	0.230	0.039	0.025	96.5%	93.5%	91.5%
$\bar{Y}_1 = 20.4$	MAR_X	Syn1_srs	0.1%	0.0%	0.9%	0.266	0.042	0.025	92.5%	95.5%	91.5%
$P_{Y_3} = 0.609$		Syn2_srs	0.1%	0.1%	0.5%	0.266	0.042	0.025	94.0%	96.0%	93.5%
$P_{Y_4} = 0.117$		SRS	4.4%	9.2%	54.0%	0.912	0.067	0.071	6.5%	56.0%	34.5%
		FX_APR	0.1%	1.2%	55.3%	0.288	0.037	0.074	93.5%	95.5%	55.0%
		RE_APR	0.0%	0.7%	5.1%	0.239	0.038	0.022	97.5%	95.5%	87.0%
		Syn1_srs	1.5%	3.3%	15.0%	0.401	0.045	0.033	77.5%	91.5%	88.0%
	MAR_X,W	Syn2_srs	1.5%	3.2%	15.0%	0.400	0.045	0.033	82.0%	94.5%	91.5%
		Syn1_lwt	0.1%	0.2%	0.9%	0.281	0.042	0.025	89.5%	93.0%	91.0%
		Syn2_lwt	0.0%	0.1%	1.2%	0.278	0.043	0.025	93.5%	95.5%	92.5%

Table 3. Comparison of RelBias, RMSE and 95% CI coverage rates for the regression coefficients of Y1, Y3 and Y4 on Y2, Population true value: $\beta_{1,Y_1}|Y_2 = 0.488$, $\beta_{1,Y_3}|Y_2 = 0.227$, $\beta_{1,Y_4}|Y_2 = 0.083$

			RelBias			RMSE			95% CI coverage		
Sampling scheme	Missingness mechanism	Methods	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$
	Complete data	Actual	–	–	–	0.103	0.065	0.098	98.0%	96.0%	90.0%
		Syn1BD	0.4%	1.1%	1.9%	0.104	0.067	0.098	96.0%	93.5%	88.0%
		Syn2BD	0.2%	2.8%	5.0%	0.103	0.067	0.100	98.0%	97.5%	91.5%
$f_2 \propto$ const.		SRS	4.6%	4.6%	24.7%	0.110	0.071	0.100	93.0%	90.0%	91.0%
	MAR_X	FX_APR	0.2%	1.0%	44.7%	0.103	0.063	0.087	97.0%	97.0%	92.5%
Actual samples BD:		RE_APR	0.3%	2.1%	22.6%	0.100	0.056	0.068	98.0%	95.5%	95.0%
$\beta_{1,Y_1\|Y_2} = 0.481$		Syn1_srs	0.0%	0.5%	2.8%	0.114	0.079	0.111	95.5%	93.0%	88.0%
$\beta_{1,Y_3\|Y_2} = 0.232$		Syn2_srs	0.2%	3.0%	4.4%	0.115	0.082	0.111	96.5%	96.5%	94.5%
$\beta_{1,Y_4\|Y_2} = 0.086$	MAR_X,W	SRS	7.3%	7.5%	45.6%	0.121	0.070	0.100	93.0%	90.5%	87.0%
		FX_APR	0.4%	1.7%	53.5%	0.114	0.064	0.087	96.5%	96.0%	91.5%
		RE_APR	0.2%	6.5%	22.9%	0.105	0.054	0.073	97.5%	96.0%	96.0%
		Syn1_srs	3.6%	2.7%	9.7%	0.123	0.076	0.105	94.5%	91.5%	91.0%
		Syn2_srs	3.5%	0.5%	4.6%	0.121	0.076	0.107	96.5%	96.0%	93.0%
		Syn1_lwt	1.8%	1.4%	2.8%	0.121	0.075	0.104	95.5%	93.0%	90.0%
		Syn2_lwt	2.2%	1.5%	2.1%	0.120	0.075	0.106	96.5%	96.0%	96.5%
$f_2 \propto h(S_i)$	Complete data	Actual	–	–	–	0.108	0.066	0.088	95.0%	96.0%	95.0%
		Syn1BD	0.1%	0.6%	2.2%	0.109	0.068	0.089	95.0%	95.0%	93.0%
		Syn2BD	0.4%	2.9%	6.5%	0.109	0.069	0.090	95.0%	96.5%	96.0%
Actual samples BD:	MAR_X	SRS	12.8%	9.1%	52.0%	0.136	0.074	0.096	89.5%	90.0%	88.0%
$\beta_{1,Y_1\|Y_2} = 0.481$		FX_APR	0.5%	0.6%	43.5%	0.114	0.069	0.079	93.5%	95.0%	97.0%
$\beta_{1,Y_3\|Y_2} = 0.229$		RE_APR	0.8%	2.5%	19.0%	0.111	0.061	0.065	95.0%	95.5%	97.0%
$\beta_{1,Y_4\|Y_2} = 0.090$		Syn1_srs	0.4%	0.7%	5.6%	0.126	0.082	0.097	94.0%	92.0%	91.5%
		Syn2_srs	0.0%	3.5%	2.7%	0.124	0.082	0.098	95.0%	94.0%	96.5%
	MAR_X,W	SRS	17.6%	12.4%	69.5%	0.141	0.069	0.101	86.0%	94.0%	83.0%
		FX_APR	0.4%	5.7%	42.2%	0.118	0.066	0.082	93.5%	95.5%	55.0%
		RE_APR	1.7%	3.1%	30.1%	0.111	0.054	0.073	97.5%	98.0%	97.5%
		Syn1_srs	6.7%	3.1%	23.0%	0.136	0.073	0.093	93.0%	94.0%	94.5%
		Syn2_srs	7.4%	0.4%	19.0%	0.136	0.075	0.095	96.0%	97.5%	97.0%
		Syn1_lwt	0.9%	0.9%	6.0%	0.130	0.075	0.092	93.0%	95.5%	93.5%
		Syn2_lwt	1.7%	2.6%	3.3%	0.126	0.076	0.094	97.0%	98.0%	97.5%

strata as if there were no strata bounds, and the stratum effects can be thought as being implicitly modeled in the random intercept term ($u_j = I_h + u_{h(j)}$).

As in the quantile estimation setting, our synthetic MI compares favorably with all of its competitors, and in most cases yields comparable results to the RE_APR for estimation of means and logistic regression parameters. In the case of rare events data, our proposed new method increases the analytical size through generating synthetic population data thus is even superior to RE_APR, consistently yielding negligible biases and close to nominal coverage. The impact of ignoring the weights in the imputation (under MAR_X,W mechanism) is less than in the quantile estimation setting, with the exception of the estimation of the continuous mean $\bar{Y}_1$, where including the weight is required to obtain approximately correct coverage.

A disadvantage of the method lies in its relative inefficiency for estimating nonlinear parameters (regression coefficients) (e.g., the synthetic MI results in unbiased point estimates but a larger RMSE than the two model-based MI methods). This is typical in that nonparametric methods cannot typically compete with their fully parametric counterparts under the correct model, and is a tradeoff made to improve robustness to model misspecification.

4. Application to NHANES III

We apply our method to the National Health and Nutrition Examination Survey (NHANES) III (1988–1994), which is designed to provide national estimates of the health and nutritional status of the civilian noninstitutionalized population of the United States aged two months and older (National Center for Health Statistics 1996). The data are obtained from a stratified, multistage area probability sampling design with oversampling of certain age and ethnicity groups. For confidentiality and computational reasons, the public-use data provides two pseudo-PSUs per stratum. Another unique feature of NHANES is that data are collected through both interview and actual physical examinations of the sampled persons. Both unit- and item-level nonresponse occurs in both components of the survey, and there is a particularly high missing rate on the body mass index (BMI) measure for youth data in the physical examination component (30%). As a popular measure of overweight status and obesity, the percentiles of BMI for children and youths are of particular interest for public health reasons. The upper percentiles and the lower percentiles are also closely monitored for overweight and underweight status, respectively. As a result, we restrict our analysis sample to children and youths from two months to 16 years of age. The Appendix provides the sample R code used to conduct the analyses below.

We estimate population quantiles (from 0.05 to 0.95 with an increment of 0.05 along with two extreme percentiles: 0.03 and 0.97) of BMI for children and youths by gender. We also estimate the proportion of such a population being covered by health insurance, overall and by race. To assure congenial inference, we include the following variables that are either of primary interest in the substantive analysis or are important predictors for BMI measures in the imputation model: age, gender, race, education, mother's BMI, father's BMI and family income (Yuan and Little 2007). We compared three different methods in our treatment of the missing data:

1) complete case analysis (CC) with design-based estimation;
2) fully parametric model-based MI using design-based estimation, within which we apply both an imputation model assuming SRS and the appropriate model conditional on all three sample design features (i.e., dummy variables indicating cluster and stratum memberships as well as the log transformation of sampling weights); and
3) our proposed finite population Bayesian bootstrap method (using SYN2 since we do not have separate weights for the first and second stages of sampling), and including the log of the weight in the imputation model.

Estimates of the median BMI and the proportion of children with health insurance are given in Table 4. The CC method appears to overestimate the median of both the BMI measure and health-insurance coverage for the full sample and race domains relative to the MI approaches, and yields the widest confidence intervals or largest standard errors as a result of decreased sample size. Then again, the median of BMI obtained from synthetic MI is quite similar to that from the model-based MI, while demonstrating some advantages in efficiency by yielding shorter intervals. The generally lower health-insurance coverage estimates under the synthetic MI relative to model-based MI might be attributable to the fact that the synthetic MI are able to capture certain interactions between the sample design variables and the regular covariate matrix which are not explicitly modeled in the fully model-based MI.

Figure 3 displays a visual comparison of the percentile estimation for the three methods under consideration. We look at how those methods perform in three different percentile ranges by gender domains: the middle percentiles from 0.5 to 0.75, the upper percentiles from 0.90 to 0.97 and the lower percentiles from 0.03 to 0.1. We chose these percentile ranges because the extreme lower and upper percentiles of BMI are typically used to monitor under- and overweight for children and youths, and there is evidence that gender difference exists in these BMI percentile ranges (particularly when age is considered, i.e., growth patterns in BMI). In general, both MI methods result in very similar BMI estimates, and they are lower than those obtained from CC analysis. This makes sense since our comparison of the distributions of age for complete cases and for missing cases on the BMI measure revealed that younger children are more susceptible to missingness, and therefore CC analysis tends to overestimate BMI by excluding those younger missing cases. The inclusion of the age variable as a predictor in the imputation model corrects such an

Table 4. Alternative methods in estimating the median of BMI and the health-insurance coverage rate, for full sample and by gender and race, respectively

		Methods		
Variable	Domain	CC	Model-based MI	Synthetic MI
BMI	Overall	17.2 [17.1, 17.4]	17.1 [16.9, 17.3]	17.0 [16.9, 17.2]
	Male	17.2 [16.9, 17.4]	17.0 [16.7, 17.2]	17.0 [16.8, 17.2]
	Female	17.3 [17.0, 17.7]	17.1 [16.8, 17.4]	17.1 [16.8, 17.3]
Health insurance	Overall	0.785 (0.020)	0.778 (0.019)	0.761 (0.019)
	White	0.822 (0.018)	0.815 (0.017)	0.799 (0.016)
	Nonwhite	0.645 (0.036)	0.643 (0.033)	0.634 (0.036)

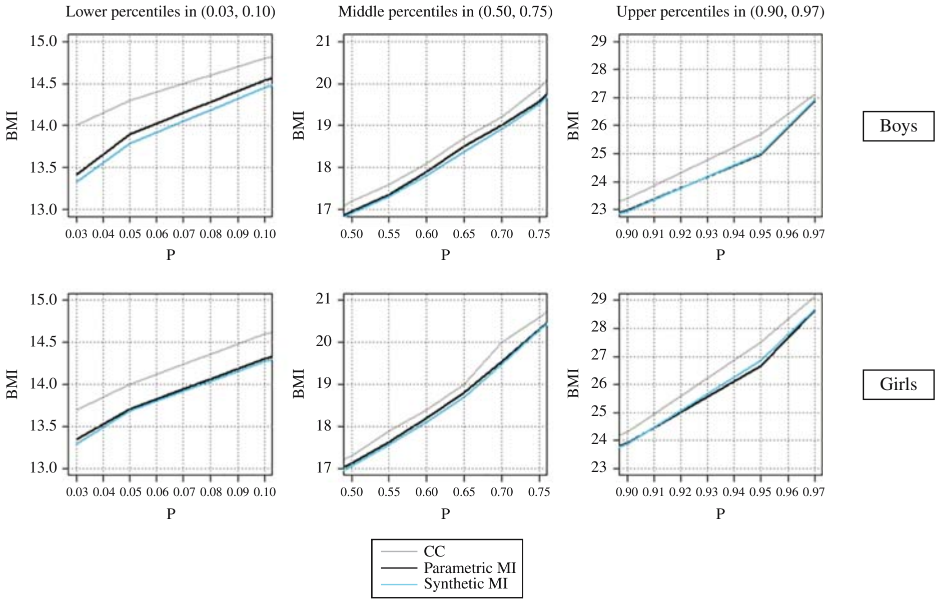


Fig. 3. Comparison of methods for quantile estimation of BMI, by gender

overestimation. The magnitude of this correction for boys is bigger than that for girls in estimating the lower percentiles (0.03, 0.05). When examining a report on BMI-for-age percentiles by gender released by the Center for Disease Control and Prevention (http://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf), we find that baby boys (corresponding to the lower quantiles here) have a relatively higher BMI, which might be at least part of the explanation.

5. Discussion

While multiple imputation has become a popular option for the analysis of missing data, some issues remain unresolved in its practical application to complex sample survey data. The complex features of sampling compounded with nonresponse in survey data often result in a rather complicated data structure, which prevents the straightforward application of the standard MI techniques (such as a multivariate normal model assuming simple random sampling). In this article, we develop a general-purpose approach to account for various design features in a highly stratified two-stage sample using a two-step synthetic MI framework. We have focused on evaluating the performance of the new method compared with existing methods with respect to several missing-data issues frequently encountered in large population-based socioeconomic and epidemiological studies. These include: i) accommodating stratification and multistage sampling in the imputation process; ii) the employment of nonstandard or non-normal imputation models for estimating probabilities of rare events; and iii) the estimation of population quantiles with multiply imputed data. (For examples that consider alternative sample designs, such as independent unequal probability of selection designs, or cluster and weighted designs without stratification, as well as estimators of quantities such as means and linear regression parameters, see Zhou (2014).

Although multiple imputation is technically valid only for maximum-likelihood estimates (Kim et al. 2006), we demonstrate that the coverage properties of the proposed method are fairly good for nonsmooth statistics. Specifically, our stratified variations of the weighted Polya posterior exhibits robustness to the loss function for estimating the upper and lower tails of the distribution function where even the appropriate model-based method (i.e., FX_APR) fails. In contrast with existing fully parametric MI methods, most of which perform poorly when applied to rare outcome binary data, the proposed method yields quite stable parameter estimates regardless of the rarity of the outcome. An alternative approach for MI estimation of quantiles that relies on estimating the CDF using a smooth regression curve is given by Wei et al. (2012), and could be used at the second-stage imputation step after the weighted finite population Bayesian bootstrap has been implemented.

It is worth stressing that our method requires only the most straightforward form of imputation modeling and combining rules for inference. This is because the effects of the complex sample design and the effect of estimating the nuisance parameters in imputation (e.g., regression parameters when the main quantity of interest is a quantile of Y) are both correctly reflected in the replication variance estimation given the design-reversed and multiply imputed synthetic populations. Any higher-level and nonlinear interactions in the covariate data, including those with the weights, clusters, or strata, will automatically be captured in the synthesizing step. However, when the imputation is conducted parametrically, as it is here, such design-variable interactions will still need to be considered if they are associated with the missingness mechanism, although the impact of misspecification will generally be attenuated. Similarly, not-missing-at-random mechanisms that are dependent on the missing values are not accommodated in this framework. Finally, we note that assuming SRS for imputation results in correct inference only at the population level: correct inference for domain estimation requires that the domains be included in the imputation model. For example, if variables X and Y are positively correlated in stratum A but negatively correlated in stratum B, this interaction will be correctly averaged over for the population inference using weighted FPBB, but if this interaction is of direct interest, it will be attenuated unless incorporated in the imputation model for the synthetic population. Further, imputing under SRS does not absolve the imputer from correctly modeling the data. To give a trivial example, assume data are sampled from two strata denoted by $Z = \{1,2\}$, where $P(Z = 1) = P(Z = 2) = .5$ in the population, and $Y|Z = 1 \sim N(5,1)$ and $Y|Z = 2 \sim N(-5,1)$, and stratum 1 is oversampled with $P(I|Z = 1) \propto .8$ The method proposed here will correct the imbalance between the strata, and assuming a two-component normal mixture model will allow imputations of Y that maintain the correct marginal distribution of Y with equal-sized components. This will allow for correct estimation of percentiles, whereas simply assuming a unimodal normal distribution will only consistently estimate the mean. Correct estimation of percentiles *within* the strata will require also conditioning on the strata, as mentioned above. We note that one advantage of the proposed method is that, with design issues cleared out of the way, more focus can be given to developing missing-data models.

We also note that the method developed here does *not* allow for the release of a small number of multiply imputed datasets to be combined using the standard Rubin rules. It *would* be possible to publically release all $L \times B \times A$ multiply imputed datasets to be analyzed using the methods developed here, although this would typically involve hundreds to thousands of datasets. Methods to allow a more modest release, with minimal impact on inference, are a topic for future research.

Future research will investigate the inferential properties of the proposed method in situations where auxiliary information on all population units is available, using a constrained version of the Polya posterior. Two other possible research directions include: (i) extending the two-step synthetic MI framework to deal with unit nonresponse problems, and (ii) extending it to deal with generating synthetic data for disclosure risk limitation.

Appendix: R Code for Using the Proposed Two-step MI Method on NHANES III

```
require(survey)
require(mice)
require(polyapost)
set.seed(seed #)

syn_bmi < -function(dt, N, Bt1, Bt2, Mt){

##Step 1: Generate synthetic populations with missing data;
#Stage 1: Create bootstrap samples from the parent sample;
          dsgn <- svydesign(ids = ~ predcl, strata = ~ pstrat, nest = TRUE, data =
          dat, weights = ~ predwt)
          dsgn.RW<-as.svrepdesign(design = dsgn, type = “subbootstrap”, replicates
          = Bt1)
          dim(dsgn.RW$repweights)
          repwt<-as.matrix(dsgn.RW$repweights)
          repwt[repwt = =0]<-NA
          dim(repwt)

          #set up arrays to hold point estimates from bootstrap samples;
          btm<-matrix(0,nrow = Bt1,ncol = 3)
          btqt<-matrix(0,nrow = Bt1,ncol = 21)
          btqtm<-matrix(0,nrow = Bt1,ncol = 21)
          btqtf<-matrix(0,nrow = Bt1,ncol = 21)


          for (j in 1:Bt1){
                  st.bb < -cbind(dat,repwt[,j])
                  #delete those units with zero weights for each bootstrap sample;
                  st.BB < -na.omit(st.bb)
                  #recode those 999 back to NA so that the mice package can be used for
                  imputation;
                  st.BB$pybmi[st.BB$pybmi = = 999] < -NA

                  #need to calculate the replicate weights;
                  Samwt < -st.BB[,9]*st.BB[,13]
                  #normalize again the adjusted weights;
                  Samwts < -Samwt*N/sum(Samwt)
                  np < -nrow(st.BB)
```

```
                ids < -seq(np)
                ns < -N-np

##Stage 2: Create unweighted synthetic populations within each bootstrap sample;
#Set up arrays to hold point estimates from imputed unweighted synthetic populations;
        fbm < -matrix(0,nrow = Bt2,ncol = 3)
        fbqt < -matrix(0,nrow = Bt2,ncol = 21)
        fbqtm < -matrix(0,nrow = Bt2,ncol = 21)
        fbqtf < -matrix(0,nrow = Bt2,ncol = 21)

        for(boott in 1:Bt2){
                l < -vector()
                smp < -wtpolyap(ids, Samwts, ns)
                #input the adjusted weights in the weighted Polya sampling algorithm;
                for (k in 1:np){
                l < -c(l,length(smp[smp = = k]))
                }
            #check if the vector of l sums up to the number of synthetic population size;
            sum(l);

            predY1 < -c(rep(st.BB[,1],l)) #bmi
            predY2 < -c(rep(st.BB[,2],l)) #race
            predY3 < -c(rep(st.BB[,3],l)) #gender
            predY4 < -c(rep(st.BB[,4],l)) #income
            predY5 < -c(rep(st.BB[,5],l)) #education
            predY6 < -c(rep(st.BB[,6],l)) #mother's bmi
            predY7 < -c(rep(st.BB[,7],l)) #father's bmi
            predY8 < -c(rep(st.BB[,8],l)) #age
            predwt1 < -c(rep(st.BB[,9],l))
            predlwt < -log(predwt1) #log of sample weight
            predCID < -c(rep(st.BB[,12],l)) #cluster ID
            predSTID < -c(rep(st.BB[,11],l)) #stratum ID

##Step 2: Multiple imputation of the unweighted synthetic populations;

        #use the imputation model including log of weight as a predictor (syn_lwt);
        temp1 < -data.frame(cbind(predY1, predY2, predY3, predY4, predY5, predY6,
        predY7, predY8, predlwt))
        temp1_imp < -mice(temp1,method = "norm", m = Mt)
        ml < -complete(temp1_imp, 'long')
        ml$bmit < -exp(ml$predY1) #back transform bmi to its normal scale
        mlmale < -subset(ml, predY3 = = 1)
        mlfem < -subset(ml, predY3 = = 2)
        multm < -cbind(as.vector(by(ml$bmit,ml$.imp,mean)),
        as.vector(by(mlmale$bmit,mlmale$.imp,mean)),
        as.vector(by(mlfem$bmit,mlfem$.imp,mean)))
```

```
     multqt < -sapply(with(ml,by(ml,.imp,function(x)quantile(x$bmit,
     c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
     multqtm < -sapply(with(mlmale,by(mlmale,.imp,function(x)quantile(x$bmit,
     c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
     multqtf < -sapply(with(mlfem,by(mlfem,.imp,function(x)quantile(x$bmit,
     c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
        fbm[boott,] < -t(apply(multm,2,mean))
        fbqt[boott,] < -t(apply(multqt,1,mean))
        fbqtm[boott,] < -t(apply(multqtm,1,mean))
        fbqtf[boott,] < -t(apply(multqtf,1,mean))
        print(boott)
        }

    btm[j,] < -t(apply(fbm,2,mean))
    btqt[j,] < -t(apply(fbqt,2,mean))
    btqtm[j,] < -t(apply(fbqtm,2,mean))
    btqtf[j,] < -t(apply(fbqtf,2,mean))
    print(j)
    }

    smpm < -apply(btm,2,mean)
    smpv < -(1 + 1/Bt1)*apply(btm,2,var)
    smpse < -sqrt(smpv)
    smpqt < -apply(btqt,2,mean)
    smpqtv < -(1 + 1/Bt1)*apply(btqt,2,var)
    smpqtse < -sqrt(smpqtv)
    smpqtm < - apply(btqtm,2,mean)
    smpqtvm < -(1 + 1/Bt1)*apply(btqtm,2,var)
    smpqtsem < -sqrt(smpqtvm)
    smpqtf < -apply(btqtf,2,mean)
    smpqtvf < -(1 + 1/Bt1)*apply(btqtf,2,var)
    smpqtsef < -sqrt(smpqtvf)

tt < -cbind(smpqt,smpqtm,smpqtf,smpqtse,smpqtsem,smpqtsef)
ss < -cbind(smpm,smpse)
write.table(tt,file = “D:/Dissertation/paper3/nhanes/synbmiqt_lwt.csv”,row.
names = FALSE,sep = “,”)
write.table(ss,file = “D:/Dissertation/paper3/nhanes/synbmimn_lwt.csv”,
row.names = FALSE,sep = “,”)
}

##Example##
syn_bmi(dt = dt, N = 100000, Bt1 = 50, Bt2 = 5, Mt = 5)
dt < -read.csv(“D:/Dissertation/paper3/nhanes/synbmi.csv”)
#Set the synthetic population size about 10 times the sample size;
N < -100000
```

```
#Normalize the weights to sum up to the assumed synthetic population size;
dt[,“predwt”] < -dt[,“predwt”]*N/sum(dt[,“predwt”])
sum(dt$predwt)
#Recode the missing values to 999;
dat[is.na(dat)] < -999
```

6. References

Anderson, D. and M. Aitkin. 1985. “Variance Component Models With Binary Response: Interviewer Variability.” *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 47: 203–210.

Cohen, M. P. 1997. “The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs.” In Proceedings of the Section on Survey Research Methods, American Statistical Association (ASA), Anaheim, CA, 1997, 635–638.

Dong, Q., M.R. Elliott, and T.E. Raghunathan. 2014. “A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sample Design.” Survey Methodology 40: 29–46

Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife.” *Annals of Statistics* 7: 1–26.

Francisco, C.A. and W.A. Fuller. 1991. “Quantile Estimation With a Complex Survey Design.” *Annals of Statististics* 19: 454–469.

Kim, J.K., M.J. Brick, W.A. Fuller, and G. Kalton. 2006. “On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling.” *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 68: 509–521. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00546.x.

King, G. and L. Zeng. 2001. “Logistic Regression in Rare Events Data.” *Political Analysis* 9: 137–163.

Kovar, J.G., J.N.K. Rao, and C.F.J. Wu. 1988. “Bootstrap and Other Methods to Measure Errors in Survey Estimates.” *Canadian Journal of Statistics* 16: 25–45.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*, (2nd ed.). New York: Wiley and Sons, New York.

Little, R.J. and H. Zheng. 2007. “The Bayesian Approach to the Analysis of Finite Population Surveys.” *Bayesian Statistics* 8: 283–302.

Lo, A.Y. 1988. “A Bayesian Bootstrap for a Finite Population.” *The Annals of Statistics* 16: 1684–1695.

McCarthy, P.J., and C.B. Snowden. 1985. “*The Bootstrap and Finite Population Sampling. Vital and Health Statistics*.” Data Evaluation and Methods Research, Series 2, No. 95. Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington

Meng, X.L. 1994. “Multiple Imputation Inferences With Uncongenial Sources of Input.” *Statistical Science* 9: 538–558. Doi: http://dx.doi.org/10.1214/ss/1177010269.

National Center for Health Statistics. 1996. *Analytic And Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANES III (1988–94)*. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland. Available at: http://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf (accessed May 22, 2014)

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/10.2307/2288945.

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.

Reiter, J.P., T.E. Raghunathan, and S.K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32: 143–149.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. 1996. "Multiple Imputation After 18+Years." *Journal of the American Statistical Association* 91: 473–489. Doi: http://dx.doi.org/10.2307/2291635.

Rust, K. and J.N.K. Rao. 1996. "Variance Estimation for Complex Estimators in Sample Surveys." *Statistics in Medical Research* 5: 381–397.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schenker, N., T.E. Raghunathan, P. Chiu, D.M. Makuc, G. Zhang, and A.J. Cohen. 2006. "Multiple Imputation of Missing Income Data in the National Health Interview Survey." *Journal of the American Statistical Association* 101: 924–933. Doi: http://dx.doi.org/10.1198/016214505000001375.

Stiratelli, R., N. Laird, and J. Ware. 1984. "Random-Effects Models for Serial Observations With Binary Response." *Biometrics* 40: 961–971. Doi: http://dx.doi.org/10.2307/2531147.

Wei, Y., Y. Ma, and R.J. Carroll. 2012. "Multiple Imputation in Quantile Regression." *Biometrika* 99: 423–438. Doi: http://dx.doi.org/10.1093/biomet/ass007.

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

Woodruff, R. 1952. "Confidence Interval for Medians and Other Position Measures." *Journal of the American Statistical Association* 47: 635–646. Doi: http://dx.doi.org/10.1080/01621459.1952.10483443.

Yang, S., J.K. Kim, and D.W. Shin. 2013. "Imputation Methods for Quantile Estimation under Missing at Random." *Statistics and Its Interface* 6: 369–377.

Yuan, Y. and R.J. Little. 2007. "Parametric and Semiparametric Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples With Item Nonresponse." *Biometrics* 63: 1172–1180. Doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00816.x.

Zhao, E. and R.M. Yucel. 2009. "Performance of Sequential Imputation Method in Multilevel Applications." In Proceedings of the Section on Survey Research Methods, American Statistical Association ASA, August, Washington D.C., 2800–2810.

Zhou, H. 2014. "Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap." Unpublished PhD Thesis

Received June 2014
Revised April 2015
Accepted April 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 257–258, http://dx.doi.org/10.1515/JOS-2016-0012

Book Review

Carol House[1]

Lin, X., Genest, C., Banks, D., Molenberghs, G., Scott, D., and Wang, J. *Past, Present, and Future of Statistical Science*. 2014. Boca Raton, FL: CRC Press. ISBN 9781482204964, 646 pp., £47.59.

This volume was commissioned by the *Committee of Presidents of Statistical Societies (COPSS)* (the societies are the American Statistical Association, the Institute of Mathematical Statistics, the Statistical Society of Canada, and the Eastern and Western North American Regions of the International Biometric Society) to celebrate the Committee's 50th anniversary and the International Year of Statistics. It is a celebration of statistical science – its past, its people, and its influence on important issues of our time. It is a book of reflections and personal stories that provide insight into these past developments and the role of statisticians within the broader science community. The book contains 52 essays in which the authors speak personally about their interests, their career decisions, the barriers they faced, and their passion for their chosen field. The contributors are award winners, having received one or more of the prestigious awards sponsored by COPSS (the awards are the R. A. Fisher Lectureship, the Presidents' Award, the George W. Snedecor Award, the Elizabeth L. Scott Award, and the F. N. David Award). The preface presents the purpose behind the volume's construction: "through the contributions of a distinguished group of statisticians, this volume aims to showcase the breadth and vibrancy of statistics, to describe current challenges and new opportunities, to highlight the exciting future of statistical science, and to provide guidance for future generations of statisticians" (p. xvii). They have succeeded.

Who should read this book? I would include anyone opening a copy of the *Journal of Official Statistics* among the target readership. This is a collection of essays that is meant to be read in a nonlinear fashion. Each essay is personal – imparting some knowledge from the past and providing inspiration. Each is short and enjoyable to read. Many of the essays are directed at young researchers. This volume would thus be of particular interest to those younger individuals and of use to their educators and mentors – those individuals who are helping develop the next generation of statisticians. Others will read the essays, smile and remember their own journeys. The essays provide important and very personal links to our past as a profession. They also provide a sense of enthusiasm for working in our profession and tackling challenges and solving problems that can have real impact on society.

The book is divided into five parts. The first part is a brief overview of the 50-year history of COPSS. It provides an appropriate introduction to the overall volume, though in itself it is not exciting reading. However, I found the tables listing all recipients of the five

[1] National Academies of Sciences, Engineering, & Medicine - Committee on National Statistics, 500 5th Street NW Keck 1137 Washington District of Columbia 20001, U.S.A. Email: chouse@nas.edu

aforementioned awards interesting. For example, the R. A. Fisher Lectureship was first awarded in 1964 to Maurice S. Bartlett, whose lecture was entitled *R. A. Fisher and the Last Fifty Years of Statistical Methodology.* Florence N. David was the first recipient (1992) of the Elizabeth L. Scott Award, and a separate award in David's name was later created in 2001.

Part II, *Reminiscences and Personal Reflections on Career Paths*, tells personal stories of the various paths that brought these individuals from different backgrounds to a passionate pursuit of statistics. Brogan says that her "educational and career paths had twists and turns and were not planned in advance, but an underlying theme throughout was my strong interest and ability in mathematics and statistics" (p. 73). Lindsay confides: "I must confess that at this time I was still a long ways from being a fan of statistics. It seemed like a messy version of mathematics constructed from a variety of disconnected black boxes . . . but the seeds of change had been planted in me" (p. 85). And there were barriers. Shaffer reflects on her high-school preparation, saying: "I wanted to take four years of mathematics, but that turned out to be a problem . . . boys were automatically enrolled in mathematics in the first semester of 9th grade, and girls in a language of their choice" (p. 50).

Part III, *Perspectives on the Field and Profession,* provides insight into such things as the impact of statistical science on society and the role of statisticians in the interplay between statistics and science. Fienberg talks about the role of the statistician "in service to the nation" and the importance of practical problems. He advises readers, especially students and junior faculty, "to get engaged in the kinds of problems I'll describe, both because I'm sure you will find them interesting and also because they may lead to your own professional development and advancement" (p. 142). Hall discusses the beginnings of computer-intensive statistics, which started about the same time as his involvement in statistical research. Lin emphasizes the importance of collaboration within the science community, saying: "I appreciate more and more to be a scientist first and then a statistician . . . [and to] closely collaborate with subject-matter scientists" (p. 192).

Part IV, *Reflections on the Discipline,* contains 24 essays that collectively cover many important past developments in statistics along with challenges and opportunities into the future. For example, Berger discusses the importance of conditioning in statistics. Dunson reflects on the "past, present, and future of nonparametric Bayesian statistics . . . on the landscape, open problems and promising directions in modern big data applications" (p. 281). Prentice discusses contributions that the statistics discipline has made, and continues to make, in the area of public health research. In the shadow of the recent financial crisis, Lai discussed statistics in the new era of finance.

Part V, *Advice for the Next Generation*, provides examples of inspiration, points to the importance of working in collaboration with others, getting published, and ends with "thirteen rules for giving a really bad talk."

In summary, *Past, Present, and Future of Statistical Science* is an excellent volume that helps us connect both to our roots and our future. The essays are independent of one another and can be read in any order. It is easy to pick up the volume, select an essay, and enjoy 20 minutes of reflection. Lindsey opines on the importance of such reflection: "one aspect of academic life that has been frustrating to me is its ruthless vitality, always rushing forward, often ending up looking like a garden sadly in need of weeding. I wish there were more reflection, more respect for the past" (p. 84). This volume assists with that reflection.