JOS

Journal of Official Statistics

An International Quarterly Published by SCB, Statistics Sweden

www.jos.nu

September 2013

Journal of Official Statistics
vol. 36, 4 (Dec 2020)

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 729–736, http://dx.doi.org/10.2478/JOS-2020-0036

Letter to the Editors

Defeating the Pandemic Requires High Quality and Ethical Official Statistics

Knowing the truth about the pandemic means that we know, for example, how many are currently infected, how many have been infected in the past and recovered, and how many have died from COVID-19. This information is essential for addressing the pandemic and its social and economic effects with appropriate government policies, rigorous work by the scientific community, and proper actions by the population, civil society and the business sector of any country. It is also essential for effective international cooperation, which is absolutely key to addressing the pandemic and its effects.

Getting as close as possible to the truth about the pandemic requires objective and impartial provision of data on COVID-19 that follows the highest quality principles in the collection, processing and dissemination of these data. A priori, there is no better candidate than official statistics to provide such data, and show us all the most accurate possible, unadulterated picture of the pandemic.

Official statistics are statistics that are developed, produced and disseminated as a public good by the National Statistics Office (NSO) and other agencies clearly identified in law as part of the National Statistical System and that aim to serve the statistical information needs of the entire society. Official statistics are provided regularly to the various branches of government, the state administration, the citizenry, the general public, the markets, the research community and the international partners of the country, as well as the international organizations of which the country is a member. Official statistics are to comply with international and national compilation standards and statistical ethics. These statistics should be our main source of data on the current pandemic, as well as other epidemics and health conditions.

This preference for official statistics as the source of truth about the pandemic is present because official statistics are a *global* public good (Georgiou 2017), providing the same information simultaneously to all users anywhere in the world. This global public good (official statistics) is needed to assist in the production of another global public good – the fight against epidemic diseases, whose urgency has been in bold relief in 2020. In addition, and very importantly, official statistics are supposed to be produced and disseminated on the basis of international *statistical principles and ethics*. These are enshrined in codes of practice and lists of fundamental principles adopted by the United Nations, and form part of national and regional legal frameworks. If – and this is an important 'if' – official statistics on the pandemic are produced and disseminated on the basis of international statistical principles and ethics that ensure the integrity and high quality of these statistics, it will bring us as close as possible to the truth about the pandemic.

Acknowledgments: The author is grateful to Hallgrimur Snorrason, Jean-Michel Durr and Gerry O Hanlon for comments on an early draft. The views expressed herein and any errors are those of the author.

So, official statistics have, *in principle,* a deontological, legal and institutional 'head-start' as a candidate in the production of statistics on the pandemic, that is, statistics on the incidence, prevalence and mortality of the pandemic. To ensure that they *in practice* end up living up to expectations about quality and ethics in this important task, such official statistics *must* adhere to the following principles (which, inter alia, draw on codifications of principles of official statistics, such as the European Statistics Code of Practice (Eurostat 2017) and the UN Fundamental Principles of Official Statistics (UNSD 2013)):

- There must be international standards and guidelines on the production and dissemination of pandemic statistics and these must be widely publicized and readily accessible to the public. The international health and epidemiology *statistics* community coordinated by World Health Organization *statisticians* must see to that,
- National official statistics on COVID-19 should strictly conform to existing international statistical standards and guidelines, by being subject to procedures to ensure that existing standard concepts, definitions, and classifications are consistently applied,
- Timeliness and periodicity of official statistics on COVID-19 should meet appropriately ambitious international release standards,
- Detailed information (metadata) about the statistical processes and primary data sources used, as well as about the statistical results on COVID-19, should be produced and disseminated and easily available to the public,
- All statistical results on COVID-19, as well as their source data and statistical processes, should be regularly assessed and validated by dedicated procedures in the official statistics agency producing them, and their quality should be regularly and openly reported to the public according to established quality criteria,
- The statistics departments of international and regional organizations dedicated to the fight against pandemics (or, alternatively, international and regional statistical organizations using relevant expertise) should regularly and rigorously assess the quality of national official statistics on COVID-19 and publish these assessments,
- The statistics departments of international and regional organizations dedicated to the fight against pandemics should reproduce, in their various publications, the national official statistics on COVID-19 (only if they can do so) with full metadata and quality information,
- COVID-19 statistics should be comparable over time and across countries. These statistics and the corresponding metadata should be presented in a form that facilitates proper interpretation and meaningful comparisons across time and across countries and regions,
- All errors discovered in published COVID-19 statistics should be corrected at the earliest possible date and publicized,
- Advance and widely publicized notice should be given regarding major revisions or changes in the methodology for producing and releasing COVID-19 official statistics. Revisions of COVID-19 data should follow standard and transparent procedures,
- Individual data (i.e., of statistical units) collected by official statistics agencies on cases of COVID-19 should be kept strictly confidential and used *exclusively* for statistical purposes, with no exceptions,

- COVID-19 official statistics should be compiled on an objective basis determined *solely* by statistical considerations. So, for example, choices of data sources and statistical methods, as well as decisions on the dissemination of these statistics, should be based solely on statistical considerations,
- All users should have equal access, at the same time, to statistical releases on COVID-19, and statistical releases and statements made in press conferences should be objective and non-partisan, and
- Finally, official statistics producers should independently–but within the confines of international standards and guidelines–decide on, and be accountable for, the methods, standards and procedures to be used in the production of COVID-19 statistics and on the time and content of statistical releases on COVID-19.

We do not aim here to assess whether any official statistics on the pandemic that may already be produced conform to the above principles. If they are going to be official statistics compiled with high quality and unimpeachable ethics, in our view they will *need to* conform to these principles. Similarly, we do not intend to pass judgment on whether international, regional and national agencies responsible for COVID-19 statistics – regarding common standards and guidelines, production, dissemination – have performed according to what is appropriate, as outlined above. They *ought to* live up to these principles. We invite observers to examine and question any available official statistics that are currently produced or may be produced in the future as to whether they meet the above list of principles and make up their own mind about the trust to be accorded to the statistics. Here, we offer only the cautionary remark that there may be room for improvement in various cases.

Political leaders and parties in office (at national or local levels) facing reelection, governments trying to attract foreign visitors, and authoritarian regimes in power needing to legitimize their hold on power might all be interested in COVID-19 statistics showing that they have ‘conquered’ the pandemic in their own jurisdictions with their epidemic mitigation efforts. At the same time, opposition politicians might also want to feed their own political narratives of alleged incompetence and negligence of the political party in power in handling the pandemic. Powerful business interests might also have ‘preferred’ narratives to be potentially supported by COVID-19 statistics. Even governments and pharmaceutical companies associated with certain therapeutic drug and vaccination initiatives may have pictures of reality they would prefer to see emerge from COVID-19 statistics. Thus, all or some of the above actors may be tempted to try to influence the production of statistics regarding the prevalence and mortality of the pandemic in their countries’ and jurisdictions’ populations. Or, at least, there may be a perception among the public that such influence is taking place. Thus, it is truly imperative that official statistics compiled on COVID-19 meet the above principles regarding ethics and quality. In this way, being of unimpeachable integrity and high reliability, these trusted statistics can serve as a bulwark against politically serving ‘fake news’ and ‘alternative facts’.

It should be noted that to respond adequately to the pandemic, the statistics on the spread and mortality of the disease itself referred to above would have to be supplemented by other statistics on the many different effects of the pandemic on social, economic, environmental and other health conditions in society. These statistics, whether well-established statistics or new flash estimates and experimental statistics, should adhere to similar principles as those outlined above.

The COVID-19 pandemic brings forth the need to think in different ways about some statistical methods and processes in identifying cases of infection with the virus. Up to now, the cases identified in each country and jurisdiction are those identified (confirmed) by medical professionals when they encounter a patient whom they assess one way or another to have the COVID-19 infection. These medical front line personnel are then supposed to report these findings up through the administrative (usually health) apparatus of the country or other administrative jurisdiction. This makes sense if the statistics to be compiled are reporting 'identified cases of COVID-19', but it does not make sense if one wants to know the total number of infected persons in a country or jurisdiction within it. In this sense, we agree with the recent letter to the JOS editors (Di Gennaro Splendore 2020, 230), which states that "figures produced by the health authorities cannot provide crucial information".

Certainly, some would argue that the present practices are good enough. We believe this is not true. To see that there is a need for official statistics on the pandemic to be produced in a different way, one can consider the following: the number of cases of COVID-19 identified (confirmed) by health authorities could give a hint about the magnitude of the problem of infection, but undoubtedly that number is a function of the number of examinations/tests that are carried out by the health system of the country or jurisdiction. One can think of situations in which test availability is intentionally or unintentionally limited, or the medical system for various reasons does not report all identified COVID-19 cases, or the population and especially certain groups in it are discouraged by culture, economics or politics, or even physical incapacity, from seeking a test or examination by a health professional when they have COVID-19-like symptoms. In addition, it is well-known that many individuals infected with COVID-19 are fully asymptomatic and are likely to never present themselves for testing or examination. In addition, testing for COVID-19 most often takes place at the initiative of individuals or businesses for their own information, or of authorities targeting certain localities/demographics. Thus, results of tests and examinations for COVID-19, as they are carried out and reported now, extrapolated to the entire population, would likely give rise to *biased* estimates of the overall incidence of COVID-19 infection in the population, as well as in certain demographic groups. Thus, the number of identified COVID-19 cases and its change over time does not provide reliable information from which to adequately estimate either the true number of currently infected individuals in the population (and the true number of new infections in a specific time period) or the true number of those that have had COVID-19 and have since recovered.

Persisting with identified (confirmed) cases of infection as the main indicator of the spread of the pandemic also sets the statistics up for criticism of credibility and for exploitation in the successful peddling of narratives that may be driven by political and economic interests, rather than by a desire to find the truth and take action on that basis.

What is needed, instead, is *an expansion of the usual tools of official statistics to include testing of biological samples from respondents in the context of a statistical survey aimed at inferring the prevalence of COVID-19 in the population*. Specifically, there is a need for proper stratified multistage sampling, on a very regularly recurring basis, of the population of the country, so that it could be tested for current COVID-19 infection (via testing for the virus) and for past infection (via testing for antibodies for the virus). Then, given appropriate size sampling and thoughtful sample design, statistical estimates of (inferences about) current and past infection of the entire population of a country and for various subpopulations can be generated. Subpopulations could consist of subnational jurisdictions (states, provinces, counties, etc.) and of specific demographic groups (by sex, age, ethnic background, etc.). The sample design of the statistical survey would have to be carefully configured to collect samples of what might be a relatively small – compared to the total – population of currently or previously infected individuals occurring in clusters.

The essential feature in what we propose is that the official statistics survey is not conducted simply by asking the respondents questions about their health status and medically related behavior and social conditions, although that should also be part of the survey, but that the respondents provide to appropriately trained survey field interviewers biomarker samples from their respiratory system (e.g., via swabbing) and blood for testing. The use of biomarkers in surveys is not new. For example, the Demographic and Health Surveys (DHS) Program has been collecting biomarker samples from interviewees in a number of surveys around the world; DHS surveys in over 50 countries have included about 20 biomarkers, relating to "a wide range of health conditions, including infectious and sexually transmitted infections, chronic illnesses (such as diabetes, micronutrient deficiencies), and exposure to environmental toxins" (DHS Program 2020).

The official statistics survey on COVID-19 we are advocating incorporates the feature of biomarker testing in appropriately sampling the population. It is important to highlight that, with this approach, there is no need to test everyone in the population to have an unbiased estimate with adequate precision (assuming an appropriate size sample) of the total number of people currently infected and the total number of people who have been infected so far. Such a survey would not be very cheap, especially as it would need to be undertaken on a very regular basis. However, its cost would be microscopic compared with the unnecessary costs – in terms of human lives lost and economic and social costs – of a pandemic, on account of not knowing the basic facts the proposed statistics would divulge to us. Policy makers should not hesitate to provide the resources needed to official statistics producers.

The proposal for such a methodology for producing COVID-19 statistics may strike some as somewhat intrusive and may raise the prospect of risks to privacy. This is actually true. This is an additional reason why statistics about COVID-19 should be compiled by official statistics producers fully committed to statistical principles and ethics and, specifically, to the principle of statistical confidentiality. As the UN Fundamental Principles of Official Statistics require: "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes" (UNSD 2013). This means that if a respondent in the survey is found to be positive for COVID-19, the information about the specific person will definitely *not* be made available to health officials, the police, administrators involved in tracking cases or enforcing isolation, or any other public agency or private entity. (The information would be made available by the official statistics producer only to the person undertaking the testing/survey.) This approach would help ensure an adequate response rate to the survey and its success in generating unbiased estimates, not only the first time it is carried out but also when it is carried out periodically, as it should be.

Statistical estimates of the total number of people currently infected by COVID-19 and the total number of people who have been infected so far would have to be supplemented by reliable and timely vital statistics. Vital statistics about deaths from COVID-19 infection would be particularly important. In general, it would be important for statistics about deaths (from any cause) to be highly reliable and timely, as they would offer a reasonable indirect measure of the toll of the pandemic in the form of 'excess deaths' measures. Such vital statistics should also be produced as official statistics (notwithstanding that the primary source would be administrative records) adhering to the principles discussed earlier in this letter. These pieces of information would then allow health scientists to have in different points in time reliable estimates of epidemiological variables, such as the susceptible part of the population, the infected part of the population and the removed part of the population (recovered or diseased). (A description of such components of the SIR (susceptibles, infected, and removed) epidemiological model can be found in Ball 2020.) This information would help estimate the degree of infectiousness and the infected fatality rate, which are important components of epidemiological models. Thus, reliable official statistics would greatly help health scientists calibrate (repeatedly over time) their models of the pandemic and, thus, improve their projections about the development of the pandemic.

The idea of the need for random sampling to understand what is happening in the population regarding the pandemic has also been advanced in a previous letter to the JOS editors (Di Gennaro Splendore 2020) and by a number of other authors elsewhere (for example, Alleva et al. 2020; Cochran 2020; Cook and Gray 2020; Trewin 2020). In this letter, our general argument for sampling the population using biomarker testing to generate statistical estimates of current and past infection of the entire population is in congruence with other authors' ideas.

In this letter, we also specifically and explicitly argue that in order for any (such) statistical estimates on the pandemic to be of adequate quality, they have to be produced as official statistics and not, in any sense, as byproducts of public health policy making. This means they can be produced by the NSO or another official statistics producer that operates in full conformity of statistical principles such as the ones outlined earlier, but they should not be produced by an institution that does not meet those principles, nor should they be produced by some mixed team of institutions/officials, of which some are not official statistics producers.

In this sense, we would argue against any proposals or statements that could be interpreted as accepting of a merging of the efforts of national statistical offices and of public health policy institutions. In such usually well-meaning visions of official statisticians 'working together' with or under officials from the public health policy/national health system side lurk significant risks to the integrity and quality of the resulting official statistics (Georgiou 2019), as some important principles, from the list provided in the beginning of this letter, would be at risk of not being met.

This does not mean that official statistics producers of COVID-19 statistics would not utilize knowledge and expertise that may exist outside their institutions; they ought to, but it would have to be done on the terms of the official statistics producer acting in independence and fully observing the statistical principles and ethics noted in this letter. In addition, our proposed approach does not mean that official statistics producers would not communicate with users of their statistics about their findings; they would, but do so fully adhering to the statistical principles above, and in particular those concerning impartiality and objectivity, as well as statistical confidentiality. There should be an exchange of knowledge and communication between official statisticians on the one hand and health scientists, providers and policy officials on the other hand, but their roles would have to be kept clear and distinct.

Thus, our proposed approach explicitly and unambiguously *disengages* the health provision, public health monitoring and policy response work on the pandemic from official statistical work on the pandemic. In order to ensure that statistics on the COVID-19 pandemic – and any future epidemic – do not fall prey to actual or suspected *conflicts of interest* or engender risks to *privacy*, the production of these statistics should be left to independent official statistics producers without the involvement of other interested parties.

Generating statistics on the pandemic in the way described above is, we believe, the best way to serve the work of health providers, public health officials, and policy makers alike. It is also the best way to serve the interests of society as a whole and of the global community.

Such official statistics can provide reliable information on the extent and the rate of the spread of the virus in a country (as well as in its individual geographic or administrative subdivisions and in its demographic groups), the true mortality rate, the extent to which herd immunity might be developing, and more. This is, of course, indispensable evidence on which to responsibly and rationally base all decisions, not only on public health policy, but also on economic policy, a multitude of social policies, national security and foreign policy, as well as international cooperation and coordination. And the same holds for decisions that have to be made by scientific researchers, families and individuals, businesses and the markets. Finally, this evidence is necessary to keep the policy makers and leaders in charge accountable during the challenging time of the pandemic. How else would one judge whether the policy responses are appropriate and successful?

In conclusion, official statistics need to 'step up to the plate' all around the world to quickly defeat the pandemic for good and minimize loss of life and economic and social distress. To do so, official statistics on COVID-19 must be produced with high quality, which requires innovation as well as the application of official statistics principles. Such approaches should be adopted not only for the current pandemic, but be part of the 'armor' in place to protect against similar attacks in the future.

References

Alleva, G., G. Arbia, P.D. Falorsi, G. Pellegrini, and A. Zuliani. 2020. "A sample design for reliable estimates of the SARS-CoV-2 epidemic's parameters. Calling for a protocol using panel data." Available at: https://web.uniroma1.it/memotef/sites/default/files/Proposal.pdf (accessed October 2020).

Ball, P. 2020. "How do epidemiologists know how many people will get Covid-19?" *Significance*. Available at: https://www.significancemagazine.com/science/648-how-do-epidemiologists-know-how-many-people-will-get-covid-19 (accessed October 2020).

Cochran, J. 2020. "Why we need more coronavirus tests then we think we need." *Significance*. Available at: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/1740-9713.01398 (accessed October 2020).

Cook, L., and A. Gray. 2020. "Official statistics in the search for solutions for living with COVID-19 and its consequences." *Statistical Journal of the IAOS* 36 (2): 253–278. DOI: 10.3233/SJI-200671. Available at: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200671 (accessed October 2020).

DHS Program. 2020. Website. Available at: https://dhsprogram.com/What-We-Do/Biomarkers.cfm (accessed October 2020).

Di Gennaro Splendore, L. 2020. "COVID-19: Unprecedented Situation, Unprecedented Official Statistics." *Journal of Official Statistics* 36 (2): 229–235. DOI: https://doi.org/10.2478/jos-2020-0012..

Eurostat. 2017. "Eurostat European Statistics Code of Practice". Website. Available at: https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142 (accessed October 2020).

Georgiou, A. 2017. "Towards a global system of monitoring the implementation of UN fundamental principles in national official statistics." *Statistical Journal of the IAOS* 33 (2): 387–397. DOI: https://doi.org/10.3233/sji-160335.

Georgiou, A. 2019. "Extracting statistical offices from policy-making bodies to buttress official statistical production." *Journal of Official Statistics* 35 (1): 1–8. DOI: https://doi.org/10.2478/jos-2019-0001.

Trewin, D. 2020. "Using Random Sampling to Learn About Covid-19 Known Unknowns." Comment in the discussion: Official Statistics in the context of the COVID-19 crisis. *Statistical Journal of the IAOS*, discussion platform. Available at: https://officialstatistics.com/news-blog/crises-politics-and-statistics (accessed 29 October 2020).

UNSD. 2013. "Fundamental Principles of Official Statistics". United Nations Statistics Division (UNSD) website. Available at: https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx (accessed October 2020).

Andreas V Georgiou

Visiting Lecturer and Visiting Scholar, Amherst College,
Former President (2010–2015),
Hellenic Statistical Authority, Greece
Amherst College, Converse Hall,
220 South Pleasant Street, Amherst, U.S.A.
Email: avgeorgiou83@amherst.edu

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 737–761, http://dx.doi.org/10.2478/JOS-2020-0037

Basic Statistics of Jevons and Carli Indices under the GBM Price Model

Jacek Białek[1]

Most countries use either the Jevons or Carli index for the calculation of their Consumer Price Index (CPI) at the lowest (elementary) level of aggregation. The choice of the elementary formula for inflation measurement does matter and the effect of the change of the index formula was estimated by the Bureau of Labor Statistics (2001). It has been shown in the literature that the difference between the Carli index and the Jevons index is bounded from below by the variance of the price relatives. In this article, we extend this result, comparing expected values and variances of these sample indices under the assumption that prices are described by a geometric Brownian motion (GBM). We provide formulas for their biases, variances and mean-squared errors.

Key words: Consumer price index; geometric Brownian motion; Jevons index; Carli index.

1. Introduction

Elementary price indices are used in inflation measurement at the lowest level of aggregation. The choice of the elementary formula does matter. For instance, in January 1999, the index formula used at the lower level of aggregation in the US consumer price index (CPI) calculations was changed to the ratio of geometric means of prices (Silver and Heravi 2007). The effect of this change was researched by the Bureau of Labor Statistics (2001) and it turned out that the change had caused a reduction of the annual rate of increase in the CPI of approximately 0.2 percentage points. Using the Boskin et al. (1996), estimates, according to which the effect of correcting a 1.1 percentage point overstatement in the CPI on federal debt was USD 1066.6 billion reduction by 2008 (see Boskin et al. (1996), we can conclude that the previously used elementary price index would generate a cumulative additional national debt from over-indexing the federal budget of approximately USD 200 billion (USD $1066.6 \cdot 0.2/1.1$ billion) over the twelve-year period up to the mid-1990s (see also Boskin et al. 1998).

In March 2013, the UK's Office for National Statistics (ONS) started to publish a new inflation index – RPIJ. This index is identical to the Retail Price Index (RPI), except it uses a geometric mean of price relatives (known as the Jevons index) rather than an

[1] Department of Statistical Methods, University of Lodz, ul. Uniwersytecka 3, 90–137, Lodz, Poland. Email: *jacek.bialek@uni.lodz.pl*.

Acknowledgments: The article is financed by the National Science Centre in Poland (grant no. 2017/25/B/HS4/00387). The article represents the author's personal opinion and does not necessarily reflect the view of Statistics Poland. I would like to thank the anonymous reviewers for their very helpful comments and inspiring remarks.

arithmetic mean of price relatives (the Carli index). At that time, the UK Statistics Authority decided that, due to the potential upward bias in the Carli index, the old RPI would no longer be recognised as a national statistic (UK Statistics Authority 2013).

The above-mentioned decision was quite controversial. On the one hand, the new RPIJ gave a much lower rate of inflation than the RPI (the RPI gave an average inflation rate of 2.9% for the years 1998–2013, whereas the RPIJ gave 2.5% (Levell 2015)). The use of the Jevons index in the CPI generated some mistrust in the official numbers in the United Kingdom. In other words, the CPI replaced the RPI for policy purposes, since the government replaced the RPI with the CPI for the indexation of state benefits, government pensions and tax thresholds (Levell 2015). On the other hand, due to the concern of the ONS about the Carli index's sensitivity to "*price bouncing*" and knowing that the Carli index failed some other important tests from the axiomatic price index theory (*time reversal test, circularity* – see Diewert (2012)), the ONS opted not to change the RPI, but rather to produce a new index. In January 2013, the ONS stated that the Carli index "did not meet international standards" (Office for National Statistics 2013). At present time, none of the 27 European Union countries makes use of the Carli index in their national price indices. Eurostat regulations do not allow the use of the Carli index in the construction of the Member States' Harmonised Index of Consumer Prices (HICP). There has been a general trend of replacing the Carli index with the Jevons or the Dutot formulas (Evans 2012). Some countries abandoned the Carli index formula in favour of other price indices over the last few decades, for example, Canada (in 1978), Luxemburg (in 1996), Australia (in 1998), Italy (in 1999), and Switzerland (in 2000). In 1996, in the United States, the Boskin Commission recommended that a Carli-like index that was used in the US CPI should be replaced by the Jevons index (Levell 2015).

There are many papers that compare the above-mentioned unweighted price index numbers. Early contributions of Eichhorn and Voeller (1976), Dalen (1992) and Diewert (1995) provide studies of properties of elementary indices from an axiomatic point of view. The differences between elementary indices, in terms of changes in the price variances, have been considered for sample indices by using Taylor approximations (see e.g. Dalen 1992; Diewert 1995; Balk 2005 for details). There are some papers that also compare the population elementary indices (Silver and Heravi 2007). The statistical approach, in which calculated elementary indices are treated as estimators of population indices, can be found in the following papers: Balk (2005), McClelland and Reinsdorf (1999), Dorfman et al. (1999) or Greenlees (2001). For instance, McClelland and Reinsdorf (1999) draw attention to the small sample bias in the case of the sample Jevons index as an estimator of its population counterpart. Finally, some authors, for comparisons, also use axiomatic and economic approaches (Diewert 1995; Levell 2015), or a sampling approach (Balk 2005), in which differences between elementary indices are explained with respect to the sampling design. The earlier literature, using the actual data underlying the consumer price index, has shown that the differences at the elementary aggregate level between the Dutot, Carli and Jevons indices can be quite substantial (see Carruthers, et al. 1980; Dalen 1994; Schultz 1995; Moulton and Smedley 1995).

In this article, we focus on only two elementary price indices, namely we consider the Jevons and Carli formulas. It has been shown (see Hardy et al. (1934) for details concerning inequalities for elementary mean values) that the difference between the Carli index and the

Jevons index is bounded from below by the variance of the price relatives. In this article, we extend this result under the assumption that prices are described by a geometric Brownian motion (GBM). We confirm some commonly known facts (such as a negative bias of the Jevons sample price index) and we also obtain some new results, including approximations for variances and mean squared errors (MSEs) of the Jevons and Carli indices and their asymptotic behaviour. According to the author's knowledge, there is a lack of works in the literature that use stochastic models with continuous time to compare elementary indices. The advantages of this approach have been emphasised in the further part of the article (see Empirical Illustration: Case 2 and Conclusions) but please note that one of new possibilities is the potential comparison of the quality of sample price indices (as estimates) at any time points, not necessarily being directly observed. Moreover, we can predict expected values of the sample Jevons and Carli indices and their other statistical characteristics for future time moments under the assumption that the nature of price processes will not change. In particular, our approach provides the possibility of forecasting expected values or variances of these sample indices or determining the above-mentioned characteristics for compared moments in time for which we do not have direct data. To be more precise: we provide new approximations to the biases, variances and mean squared errors of these indices for the above-mentioned stochastic price process. It is shown that these approximations may strongly depend on the sample size and price volatilities, but this remark is not identical for both considered indices. For instance, it is shown that the sample Carli index is an unbiased estimator of the unweighted population parameter describing the price change and the sample Jevons index is an asymptotically unbiased estimator of the same parameter. There are some other practical conclusions that can be drawn from our research; for instance, that the expected value of the Jevons index is sensitive to price volatility only in the case of small sample sizes, while the expected value of the Carli index does not depend on price dispersion (see Tables 4 and 5).

The article is organised as follows: Section 2 presents unweighted Jevons and Carli indices. Section 3 starts from the introduction to the GBM price model and compares the above-mentioned elementary indices in this area. Section 4 presents two empirical illustrations of the previously discussed theoretical results: the first study concerns prices of tomatoes sold in some number of supermarkets, the second one considers prices of mountain bikes sold via the largest online e-commerce platform in Poland. Section 5 discusses the results from our simulation study and examines the influence of price volatility on differences between the discussed sample indices; Section 6 lists the main conclusions derived from both empirical and simulation studies.

2. Unweighted Jevons and Carli Indices

There are several elementary price indices in the literature (Von der Lippe 2007; Consumer Price Index Manual, CPI Manual 2004 chap. 20). In particular, we have the following formulas:

- the Carli price index (Carli 1804)

$$P_C = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i^t}{p_i^0}, \tag{1}$$

- and the Jevons price index (Jevons 1865)

$$P_J = \prod_{i=1}^{N} (\frac{p_i^t}{p_i^0})\frac{1}{N} = \frac{\prod_{i=1}^{N} (p_i^t)^{1/N}}{\prod_{i=1}^{N} (p_i^0)^{1/N}} = \frac{\exp[\frac{1}{N}\sum_{i=1}^{N} \ln(p_i^t)]}{\exp[\frac{1}{N}\sum_{i=1}^{N} \ln(p_i^0)]}, \tag{2}$$

where the time moment $\tau = 0$ is considered as the basis, N is the number of items observed at times 0 and t, p_i^τ denotes the price of the i-th item at time τ. The Carli index is an arithmetic mean of price relatives (partial indexes), whereas the Jevons index is a geometric mean. As a consequence, these indices satisfy the classic inequality for arithmetic and geometric means

$$P_J \leq P_C. \tag{3}$$

The difference between the Carli index and the Jevons index is bounded from below by the variance of the price relatives $D^2(p_i^t/p_i^0)$ (for details concerning inequalities for elementary mean values, see for example: *Chapter II* in Hardy et al. 1934):

$$P_C - P_J \geq D^2(\frac{p_i^t}{p_i^0}), \tag{4}$$

and thus the analogical inequality holds for their expected values. From the point of view of the axiomatic price index theory, the Jevons index seems to be better, that is, it satisfies the main tests (axioms), whereas the Carli index does not satisfy the time reversal test and circularity (Levell 2015). Price indices defined in Equations (1) and (2) can be treated as sample indices, since they are estimators of unknown real values of population indices. In particular, the sample Carli index (1), being an arithmetic mean of price relatives, is a consistent estimator of the population Carli index that can be expressed as follows:

$$I_C = E[\frac{p^t}{p^0}]. \tag{5}$$

Similarly, the sample Jevons index (2), as a ratio of the exponents of two sample means of log prices (Silver and Heravi 2007), is a consistent estimator of the following population Jevons index, that is,

$$I_J = \frac{\exp[E[\ln(p^t)]]}{\exp[E[\ln(p^0)]]}. \tag{6}$$

At this point, the reader needs a certain clarification. First of all, following the work of Silver and Heravi (2007), in the presented approach, population indices are unknown a priori values expressed by expected values of price relatives or log-prices. We consider here a homogeneous aggregate, and thus a random variable, that is, the price from the examined period or the base period, reflects the prices of items of the same product sold in a given period by different establishments. Assuming a specific probability distribution for the prices from the base and the analysed period, or, as in our article, adopting a specific stochastic process describing the price process in the whole examined period of time, we are able to determine population indices without referring to the size of the population (which in fact is difficult to determine in practice), and the determination of the parameters of this distribution or process is a technical issue (see Subsection 3.2). We want to estimate population indices using the sample indices described by Equations (1) and (2) determined on the basis of a drawn sample of N matched items. In the literature, however, one can encounter a slightly different approach in which in the case of homogeneous aggregate, the target (or population) price index is the unit value index, and the population Jevons and Carli indices are defined analogously to Equations (1) and (2), where this time N denotes the size of the population, and sample elementary indices are calculated on the basis of a random sample that constitutes a subset of the above-mentioned population (see e.g. Diewert 1995; Balk 2005). Secondly, in the quoted work of Balk (2005), the author considers two scenarios of obtaining the sample. The first scenario assumes the simple random sample drawn without replacement, which means that each element of the population has the same probability of being included in the sample. In the other scenario, the more "important" elements of the population have a larger probability of being included in the sample than the less important elements. As the author himself admits, both scenarios are "more or less representative of actual statistical practice". Nevertheless, this article implicitly assumes that the first scenario is implemented.

3. Basic Statistics of the Carli and Jevons Sample Indices in the Stochastic Model

3.1. The GBM Price Model

A geometric Brownian motion (GBM) (also known as exponential Brownian motion) is a continuous-time stochastic process in which the logarithm of the randomly varying quantity follows a Brownian motion (also called a Wiener process) with drift (see Oksendal 2003; Privault 2012). It is an important example of stochastic processes satisfying a stochastic differential Equation (SDE, see Subsection 3.2. for more details). In particular, the GBM model is used in mathematical finance to model stock prices in the Black–Scholes model (Privault 2012; Ross 2014) and, for instance, to model prices of derivatives (Hull 2018). In the literature, we can encounter many other applications of this model, for instance: modelling of unit prices of Open Pension Funds (Gajek and Kałuszka 2004; Białek 2013), modelling of oil prices (Meade 2010; Nwafor and Oyedele 2017), or modelling of electricity prices (Barlow 2002). The GBM price model can be also used for generalisations of the Divisia's approach in the price index theory (Białek 2015).

The main arguments for using the GBM price model are as follows:

(a) the expected returns (relative price changes) are independent of the value of the process (price), which is consistent with what we would expect in reality;
(b) the GBM process only assumes positive values, just like real commodity prices;
(c) the GBM process shows the same kind of 'roughness' in its paths as we see in real prices; and
(d) estimations of its parameters are relatively easy.

Nevertheless, the GBM model also has some drawbacks: (A) the price volatility is assumed constant in this model; (B) the GBM model does not take into account possible jumps of prices caused by unpredictable events or news (see the final remarks in Conclusions, Section 6) since the path in GBM model is continuous. However, due to the advantages of using the GBM price model presented above, we decided to apply it in our work. The application of more advanced models, being generalisations or extensions of the GBM model (see e.g. Kou 2002; Kühn and Neu 2008; You-Sheng and Cheng-Hsun 2011; Hong-Bae and Tae-Jun 2015), in the theory and practice of price indices is our future aim.

3.2. Comparison of Carli and Jevons Indices Under the GBM Price Model

In practice, unweighted indices (such as the Carli, Jevons or Dutot indices) are used at the lowest level of aggregation in the CPI measurement (Von der Lippe 2007). In general, the calculation of the Consumer Price Index (CPI) proceeds in two (or more) stages. In the first stage, elementary indices are estimated for the elementary expenditure aggregates of the CPI. In the second and subsequent stages of data aggregation, these elementary price indices are combined to obtain higher-level price indices using information on the expenditures on each of the elementary aggregates as weights (CPI Manual 2004 chap. 20). In other words, these higher-level price indices are aggregated further through expenditure-weighted averages into “sections” that, in turn, are aggregated into “groups” (Levell 2015).

Let us go back to the lowest level of aggregation. In the very first stage, in which the Office for National Statistics (ONS) does not have expenditure information, each matched item is observed in many monitoring points (sampled outlets) in the country. Thus, for a given homogeneous set of items, we have a corresponding set of monthly prices collected systematically for each month. As a consequence, most authors treat the above-mentioned set of prices from a given month as a realisation of one random variable and they assume a common price distribution for these prices (Silver and Heravi 2007; Levell 2015). In other words, elementary price indices concern rather a given kind of good (product) observed in many places (outlets), and thus described by a vector of different prices being realisations of the same random variable. According to the above-mentioned remark, let us assume one, common price distribution, that is, we assume that all price processes (representing the considered matched items) can be described by a geometric Brownian (Wiener) motion (GBM), also known as an exponential Brownian motion. To be more precise, we assume that the given $i-$th price process satisfies the following stochastic differential equation

$$dp_i^t = \alpha p_i^t dt + \beta p_i^t dW_i^t, \tag{7}$$

where the percentage drift α and the percentage volatility β are constant, and $\{W_i^t : 0 \leq t < \infty, i = 1, 2, ..., N\}$ are independent Wiener processes. The solution for the stochastic differential Equation (7) is as follows (Oksendal 2003; Jakubowski et al. 2003):

$$p_i^t = p_i^0 \exp((\alpha - \frac{\beta^2}{2})t + \beta W_i^t), \tag{8}$$

and we assume that all initial prices p_i^0 are deterministic. As a consequence, we obtain

$$E(P_i^t) = \exp(\alpha t), \tag{9}$$

and

$$Var(P_i^t) = \exp(2\alpha t)[\exp(\beta^2 t) - 1], \tag{10}$$

where P_i^t is a $i-$th price relative and $\exp(\alpha t)$ is the (unknown) population price index that we want to estimate (Oksendal 2003; Jakubowski et al. 2003). Let us denote by $\mu_t = E(P_i^t)$ and by $\sigma_t^2 = Var(P_i^t)$ for any value of i. Let us note that the sample Jevons index can be expressed as follows:

$$P_J = \prod_{i=1}^{N} (P_i^t)^{\frac{1}{N}} = \prod_{i=1}^{N} \exp\,(\frac{\alpha - \beta^2/2}{N} t + \frac{\beta}{N} W_i^t), \tag{11}$$

or equivalently

$$P_J = \exp\,((\sum_{i=1}^{N} \frac{\alpha}{N} - \frac{1}{2}\sum_{i=1}^{N} (\frac{\beta}{N})^2)t + \sum_{i=1}^{N} \frac{\beta}{N} W_i^t)\exp\,(\frac{1}{2}(\sum_{i=1}^{N} (\frac{\beta}{N})^2 - \sum_{i=1}^{N} \frac{\beta^2}{N})t). \tag{12}$$

Let us denote by $vol(t, \beta, N)$ a component connected with price volatilities, that is,

$$vol(t, \beta, N) = \exp\,(\frac{1}{2}(\sum_{i=1}^{N} (\frac{\beta}{N})^2 - \sum_{i=1}^{N} \frac{\beta^2}{N})t) = \exp\,(\frac{1-N}{2N^2}\beta^2 t). \tag{13}$$

We have

$$E(P_J) = vol(t, \beta, N)\prod_{i=1}^{N} E[\exp\,((\frac{\alpha}{N} - \frac{1}{2}(\frac{\beta}{N})^2)t + \frac{\beta}{N} W_i^t), \tag{14}$$

From Equations (9) and (14), we obtain the following expected values of the sample Carli and Jevons indices:

$$E(P_C) = \frac{1}{N}\sum_{i=1}^{N} E(P_i^t) = \exp\,(\alpha t) = \mu_t, \tag{15}$$

$$E(P_J) = vol(t, \beta, N)\exp\,(\frac{\alpha}{N} t)^N = vol(t, \beta, N)\mu_t. \tag{16}$$

The immediate conclusion from Equations (15) and (16) is a known fact that the sample Carli index is an unbiased estimator of the parameter μ_t and the sample Jevons index is a biased estimator of the same parameter, in which its bias is very small in practice (when the sample size N is big) and it equals

$$bias(P_J) = E(P_J - \mu_t) = (vol(t, \beta, N) - 1)\mu_t. \tag{17}$$

Moreover, from Equation (10) we obtain the variance of the sample Carli index, that is,

$$\begin{aligned} Var(P_C) &= \frac{1}{N^2} Var(\sum_{i=1}^{N} \frac{p_i^t}{p_i^0}) = \frac{1}{N^2}\sum_{i=1}^{N} Var(P_i^t) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \exp\,(2\alpha t)[\exp\,(\beta^2 t) - 1] = \qquad (18) \\ &= \frac{1}{N}\mu_t^2[\exp\,(\beta^2 t) - 1] = \frac{\sigma_t^2}{N}. \end{aligned}$$

Let us note that from (11) it holds that

$$\ln P_J = (\alpha - \frac{\beta^2}{2})t + \frac{\beta}{N}\sum_{i=1}^{N} W_i^t, \tag{19}$$

and thus, since $Var(W_i^t) = t$, we obtain

$$Var(\ln P_J) = \frac{\beta^2}{N^2}Nt = \frac{\beta^2 t}{N}. \tag{20}$$

From the Taylor's approximation rule, we know that $\ln(x) \approx x - 1$ for $x \approx 1$, and thus we obtain that $Var(\ln P_J) \approx Var(P_J)$. Let us note that in practice the value of $\beta^2 t$ will be close to zero, and thus using once again the Taylor's approximation rule (i.e., $\exp(x) - 1 \approx x$ for small values of x) we obtain

$$Var(P_i^t) = \exp(2\alpha t)[\exp(\beta^2 t) - 1] \approx N\mu_t^2 Var(P_J). \tag{21}$$

Thus, finally we obtain

$$Var(P_J) \approx \frac{\sigma_t^2}{\mu_t^2 N}. \tag{22}$$

The overall performance of an estimator can be summarised by its mean-squared error (MSE), which measures its expected squared deviation from the true population value of the parameter of interest. The MSE can be expressed as follows:

$$MSE(\hat{\theta}) = E(\theta - \hat{\theta})^2 = Var(\hat{\theta}) + bias^2(\hat{\theta}), \tag{23}$$

where $\hat{\theta}$ denotes a considered estimator of population parameter θ. In our case, from Equations (17), (18) and (22), we obtain the following values of mean-squared errors of the sample Carli and Jevons indices:

$$MSE(P_C) = \frac{\sigma_t^2}{N}, \tag{24}$$

$$MSE(P_J) \approx \frac{\sigma_t^2}{\mu_t^2 N} + (vol(t, \beta, N) - 1)^2 \mu_t^2. \tag{25}$$

Let us note that from Equation (13) we obtain for the fixed values of t and β

$$\lim_{N\to\infty} vol(t, \beta, N) = \exp(0) = 1. \tag{26}$$

In practice, the number N is large, and thus we can use the following approximation:

$$E(P_J) = vol(t, \beta, N)\mu_t \approx \mu_t. \tag{27}$$

From Equations (17) and (27), we obtain the following asymptotic property of the Jevons formula

$$\lim_{N\to\infty} bias(P_J) = \lim_{N\to\infty}(vol(t, \beta, N) - 1)\mu_t = 0, \tag{28}$$

and, for big sample sizes, we obtain

$$MSE(P_J) \approx \frac{\sigma_t^2}{\mu_t^2 N} + (vol(t, \beta, N) - 1)^2 \mu_t^2 \approx \frac{\sigma_t^2}{\mu_t^2 N}. \tag{29}$$

The immediate conclusion is that the bias of the sample Jevons index is negative (see also Greenlees 2001) and the estimator P_J is asymptotically unbiased, see Equation (28). From Equations (24) and (29), we conclude that the MSEs of the sample Carli and Jevons indices can be reduced to zero by increasing the sample size. The MSEs of these sample indices are decreasing functions of the sample size and they tend to zero if $N \rightarrow \infty$. The following example shows that it is possible that $MSE(P_c) - MSE(P_J)$ is greater or smaller than zero, but in practice, this difference will be very small. Although our results, similarly to those obtained by Levell (2015), suggest that in general the ratio of the considered MSEs is greater or smaller than 1 depending on variances of prices or indices, from Equations (24) and (29), we obtain the following approximation for big sample sizes: $MSE(P_c)/MSE(P_J) \approx \mu_t^2$. Thus, in the case of increasing prices (i.e., when $\mu_t^2 > 1$), from the large data set, we can expect that $MSE(P_c) > MSE(P_J)$ – see also our results presented in Table 3.

3.2.1. Example

Let us consider a random sample connected with the given item observed in N drawn monitoring points during the unit time interval. In other words, we observe a homogeneous group of identical items with different N prices described in Equation (8). Let us denote the unknown population parameters at the end of the unit time interval as μ and σ. We take into consideration the following values of parameters of price processes: $\mu \in [0.8, 1.2]$ and $\sigma \in [0, 0.2]$. Let us denote by $\Delta = MSE(P_c) - MSE(P_J)$, where MSEs of indices are defined in (24) and (25). Differences Δ (as functions of parameters μ and σ) calculated for $N \in \{5, 10, 50, 500\}$ are presented in Figure 1.

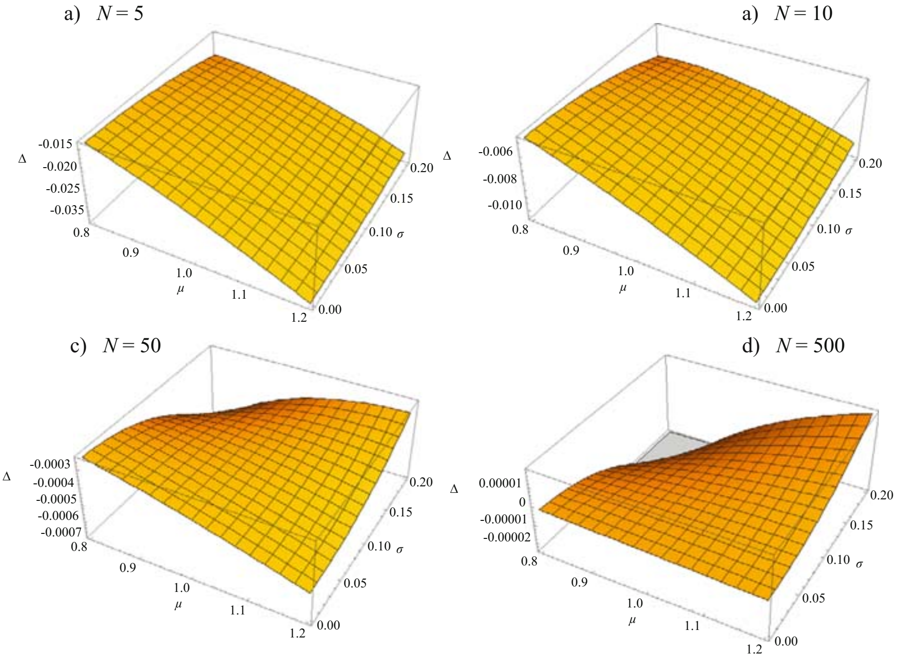


Fig. 1. Differences between MSEs of indices (Δ) depending on μ,σ and N.

Based on Figure 1, we can conclude that the difference $MSE(P_c) - MSE(P_J)$, as a function of parameters of price processes, is negative as a rule. Nevertheless, when the sample size is big and approximations Equations (27) and (29) start to work, the MSE of the sample Carli index is minimally bigger (smaller) than the MSE of the Jevons index if $\mu > 1$(respectively $\mu < 1$).

4. Empirical Illustration

4.1. Case 1

As it was mentioned above, in practice, at the lowest level of aggregation we collect prices of the considered group of items, that is, each item is observed in many monitoring (sales) points in the country. Let us suppose we observe an item number i_0 during the time interval $[0, t]$ and we have N sample prices of this item at time 0 $(p^0_{i_0k} : k = 1, 2, ..., N)$ and N sample prices of this item at time t $(p^t_{i_0k} : k = 1, 2, ..., N)$. We obtain N sample price relatives $(P^t_{i_0k} = p^t_{i_0k}/p^0_{i_0k} : k = 1, 2, ..., N)$ and, although we assume that prices are described by a geometric Brownian motion, we do not know the real values of the drift α and the volatility β in the price population. In this article, we propose using very simple estimators of α and β that are an immediate consequence of the fact that the mean of observed price relatives should be approximated by $\exp(\alpha t)$ and the standard deviation of observed price relatives should be approximated by $\exp(\alpha t)\sqrt{\exp(\beta^2 t) - 1}$ (under assumptions of the GBM model). These estimators are as follows:

$$\hat{\alpha} = \frac{\ln \bar{P}^t_{i_0}}{t}. \tag{30}$$

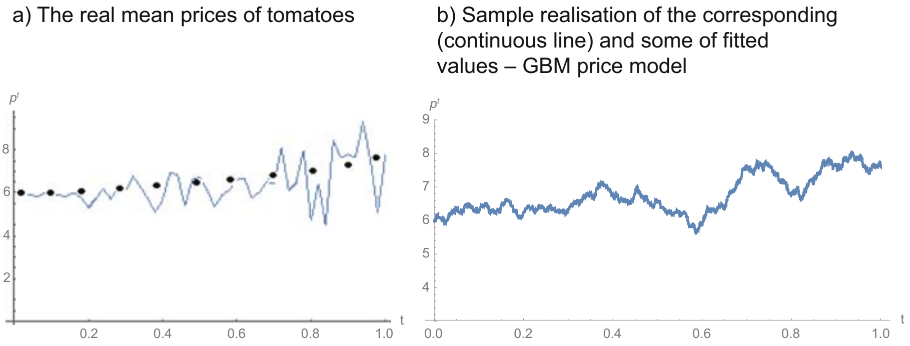
$$\hat{\beta} = \sqrt{\frac{\ln \{\exp[2(\ln S^t_{i_0} - \hat{\alpha} t)] + 1\}}{t}}. \tag{31}$$

where $\bar{P}^t_{i_0}$ and $S^t_{i_0}$ denote the arithmetic mean and the standard deviation of price relatives $(P^t_{i_0k} : k = 1, 2, ..., N)$ at time t. Estimators (30) and (31) are quite effective for the number of monitoring points over 100. Having estimated parameters α and β, we can estimate $vol(t, \beta)$ function, and thus we can approximate $bias(P_J)$ or $MSE(P_J)$ (see Equations (28) and (29)). Treating values of $\bar{P}^t_{i_0}$ and $S^t_{i_0}$ as good enough estimates of μ_t and σ_t we can approximate $Var(P_C)$, $Var(P_J)$ and $MSE(P_C)$ (see Equations (18), (22) and (24)). We apply this observation to the real data set connected with prices of 1 kg of tomatoes in Poland. We collect data about these prices from $N = 115$ Polish supermarkets observed during the interval: 01.12.2017 − 19.01.2018 (for convenience, we normalise the interval to [0, 1]).

In this experiment, we are going to estimate parameters of the GBM model using only the first and the last time moments of observations, that is, we compare only prices from 01.12.2017 and from 19.01.2018, since the interval is very short, and we use Equations (30) and (31) for estimation of these parameters. In a more practical case (see, for instance, Subsection 4.2), we take into account all time moments from the considered interval for the estimation purpose. Nevertheless, this time we want to verify whether we can still obtain good approximations of considered price indices while limiting our analysis to only two compared time moments (the base one and the current one). Our hypothesis states that

Equations (15) and (16) may approximate the real value of price indices effectively even if estimates of parameters α and β are based on only the first and the last moment of a short time interval (the only condition is a lognormal price distribution). After calculations, we obtain $\hat{\alpha} = 0.18912$ and $\hat{\beta} = 0.22629$. The real mean price process and sample realisation of the corresponding geometric Brownian motion are presented in Figure 2. We are aware of the fact that while estimating the above-mentioned parameters, none of internal time moments were used, and thus the theoretical price movements (see Figure 2b) may not be fitted to the empirical ones perfectly (Figure 2a). Nevertheless, we evaluate the fit of the GBM model to the data by determining at least simple statistics such as the Root Mean Squared Error (RMSE) or the Mean Absolute Percentage Error (MAPE). Our results ($RMSE = 0.0228$, $MAPE = 5.1938\%$) are acceptable, which may be a consequence of a lognormal distribution of price relatives. In fact, if the GBM model is a good proxy for price dynamics, then equivalently the price relatives, calculated for any fixed current time moment t, are well-described by a lognormal distribution (in our case, the Kolmogorov-Smirnov test of log-normality of price relatives p_{ik}^1/p_{ik}^0 returns p-value equals 0.112). The empirical distribution of observed price relatives $(P_{ik}^1 = p_{ik}^1/p_{ik}^0 : k = 1, 2, ..., 115)$ is presented in Figure 3. Estimates for the Carli and Jevons indices (the day of 01.12.2017 is fixed as a base time period), their sample variances and MSEs are presented in Table 1.

Our hypothesis is confirmed, that is, in the case of lognormal price relatives, Equations (15) and (16) may approximate the real value of price indices effectively, even if estimates of model parameters α and β are based on only the first and the last moment of a short time interval. Moreover, it can be observed that the difference between the observed Jevons and Carli indices is quite big due to the fact that the price volatility is large. Obviously, the sample Carli index is unbiased, while the bias of the sample Jevons index is negative and equals -0.0355. As once could expect, the bias of the Jevons index is crucial since the analysed price processes have strong fluctuations (see Equations (13) and (17)). In the analysed case, the variance and the coefficient of variation of the sample Carli price index are bigger than those calculated for the Jevons index. Please note that the precision of estimation of the expected value could be better in the case of the Jevons formula (the estimation is perfect for the Carli price index), that is, we have a theoretical value 1.2079 versus the empirical one: 1.1724. One of the reasons for this is the high price volatility, as the β parameter is above 0.2. The second reason, mentioned earlier, is that the estimation


a) The real mean prices of tomatoes

b) Sample realisation of the corresponding (continuous line) and some of fitted values – GBM price model

Fig. 2. Empirical and theoretical tomato price processes (Poland, 01.12.2017 − 19.01.2018).

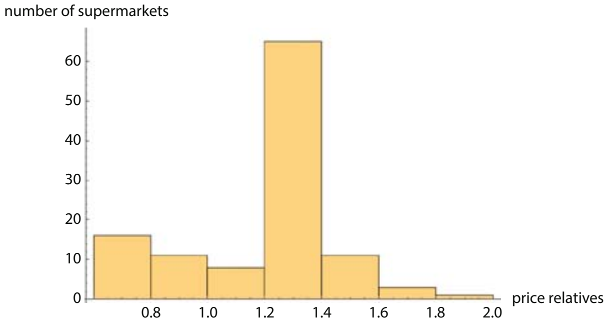


Fig. 3. The empirical distribution of observed tomato price relatives.

Table 1. Basic characteristics of the sample Jevons and Carli indices.

Characteristics	Jevons index	Carli index
Observed index value	1.1724	1.2082
Expected value (*)	1.2079	1.2082
Standard deviation (*)	0.0211	0.0258
Coefficient of variation (*)	0.0174	0.0213

(*) Values obtained under the GBM price model for $\hat{\alpha} = 0.18912$ and $\hat{\beta} = 0.22629$.

of the GBM model parameters was made only for two time moments. When we have long time series and the model fit to the data is better, then we can also expect better approximations of the characteristics of the considered indices (see Case 2). It seems, therefore, that the discussed approach and method of estimation can be used only for very short time intervals and with the log-normality of price distribution.

4.2. Case 2

Case 1 concerns a situation in which we only have price observations of an item for two compared periods. In Case 1, we assumed that the process that allowed prices to go from the known state at the moment 0 to the known (observed) state at the moment t was a geometric Brownian motion. Let us add that we made this assumption without any special possibility of its verification, or – due to only two moments of observation – without the possibility of assessing the fit of the GBM model to reality, and the estimators of the model parameters (30) and (31) were determined in a heuristic manner. Much more practical is a situation in which we have a long time series describing prices, and then we estimate the parameters α and β using the GBM model dedicated estimators (usually Maximum Likelihood Estimators), and finally, we evaluate the fit of the GBM model to the data by determining at least simple statistics such as MAE, MAPE, RMSE, and so on. Please note that every time we calculate the level of the fit of GBM models or sample indices (based on the GBM model) we compare expected values of these stochastic processes to empirical observations. If the researchers become convinced that the GBM model describes the behaviour of prices well, then they have a basis for using the theoretical results obtained in Section 3. We will demonstrate this behaviour on the example of real data regarding the prices of mountain bikes sold in 2018 via *allegro.pl*, which is one of the largest online e-commerce platforms in Poland. To be more precise: we used weekly data on mountain bikes sold from 01.01. 2018 to 31.12.2018 (48 observations), and for demonstration purposes we decided to take into consideration $N = 6$, the most popular in the considered time interval, mountain bike models with a tyre diameter of 28 inches (Denver, Davos, Indiana, Shimano, Romet, and Kross). Data were filtered and aggregated (we used unit values aggregated to one week as prices) by using a special tool provided by allegro.pl, that is, the *TradeWatch* (https://www.tradewatch.pl/index.jsf). After aggregating, we normalised each time series by dividing prices by the first observed price and we normalised the time interval of observations to [0,1], that is, we set partial indices $\tilde{p}_i^t = p_i^t / p_i^0$ and time moments $t_k = \frac{k-1}{47}$ for $k = 1, 2, ..., 48$, and thus we obtained $\tilde{p}_i^0 = 1$, $t_1 = 0$ and $t_{48} = 1$. For each normalised price process, the parameters α and β of the GBM model were estimated by using the following unbiased estimators (Privault 2012):

$$\hat{\alpha}_i = \frac{1}{T-1} \sum_{k=1}^{T-1} \frac{1}{t_{k+1} - t_k} (\frac{p_i^{t_{k+1}} - p_i^{t_k}}{p_i^{t_k}}), \tag{32}$$

$$\hat{\beta}_i = \sqrt{\frac{1}{T-2} \sum_{k=1}^{T-1} (\frac{p_i^{t_{k+1}} - p_i^{t_k}}{p_i^{t_k}} - \hat{\alpha}_i (t_{k+1} - t_k))^2}, \tag{33}$$

where $i = 1, 2, ..., 6$ and $T = 48$. We also matched the GBM model for the time series of average bike prices, for the time series $\{\bar{p}_{t_1}, \bar{p}_{t_2}, ..., \bar{p}_{t_{48}}\}$ where $\bar{p}_{t_k}$denotes the arithmetic mean of normalised prices of considered mountain bikes sold during the kth week. For the above-mentioned purpose, we used estimators with an analogical form to those defined in Equations (32) and (33), namely

$$\hat{\alpha}_T = \frac{1}{T-1} \sum_{k=1}^{T-1} \frac{1}{t_{k+1} - t_k} (\frac{\bar{p}_{t_{k+1}} - \bar{p}_{t_k}}{\bar{p}_{t_k}}), \tag{34}$$

$$\hat{\beta}_T = \sqrt{\frac{1}{T-2} \sum_{k=1}^{T-1} (\frac{\bar{p}_{t_{k+1}} - \bar{p}_{t_k}}{\bar{p}_{t_k}} - \hat{\alpha}_T (t_{k+1} - t_k))^2}, \tag{35}$$

where $k = 1, 2, ..., 48$ and $T = 48$. Justification for matching the GBM model to the mean price process can be as follows:

(1) the fixed time moment prices described by the GBM model are log-normally distributed,

(2) the sum of log-normally distributed variables can be approximated by a lognormal distribution quite effectively (see Fenton-Wilkinson (FW) approximation: Fenton 1960; Cobb et al. 2012),

(3) it is easy to prove that if X is log-normally distributed, then cX is also log-normally distributed for any positive c, in particular for $c = 1/N$.

Figure 4 presents observed (normalised) price processes and corresponding estimated price processes obtained from the matched GBM models (for each mountain bike case, information about estimated parameters α and β, as well as the level of RMSE is added below the right graph). Figure 5 presents the observed (normalised) average price process and the corresponding matched GBM model. The results obtained ($\hat{\alpha}_T = 0.1737$ and $\hat{\beta}_T = 0.006636$ with a low level of $RMSE = 0.0053$) allowed us to determine the expected values, standard deviations and coefficient of variations of the sample Carli and Jevons indices under the GBM price model (theoretical results are presented in Section 3). The

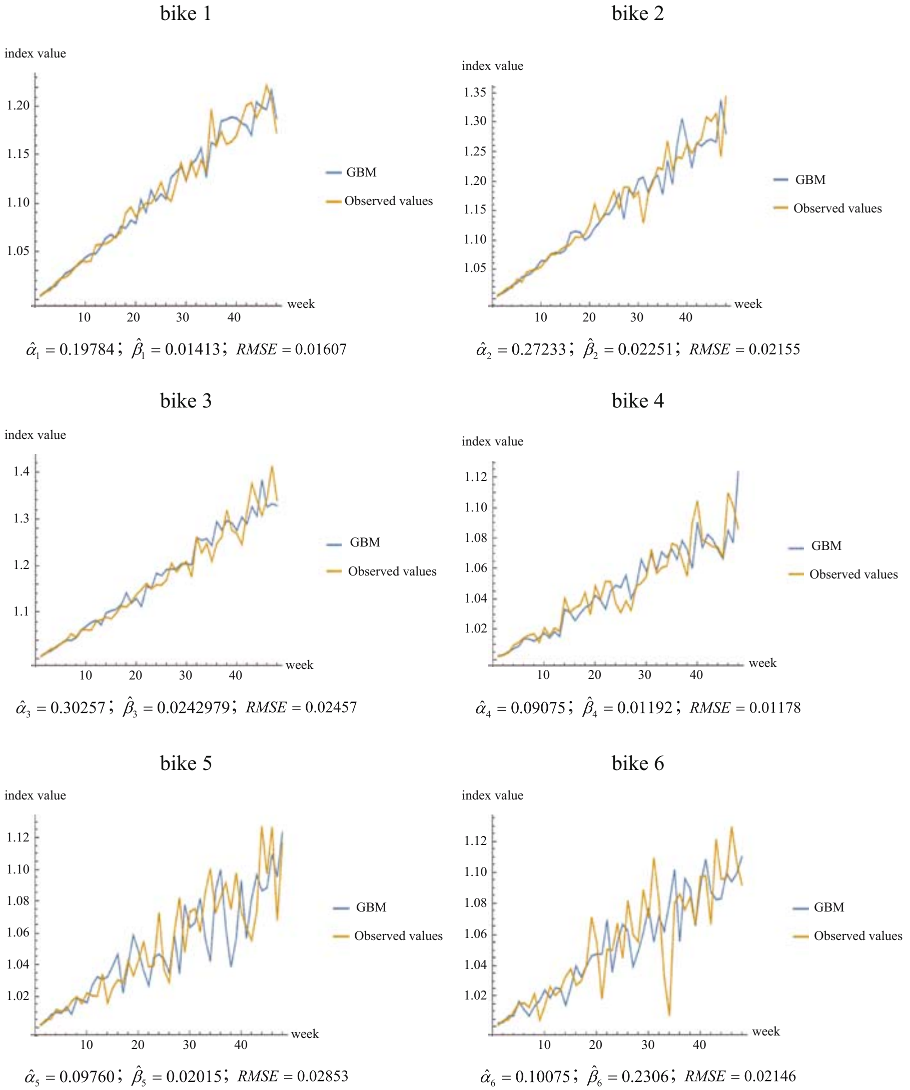


Fig. 4. Observed normalised bike prices (Observed values) and sample realisations of matched GBM price models (GBM).

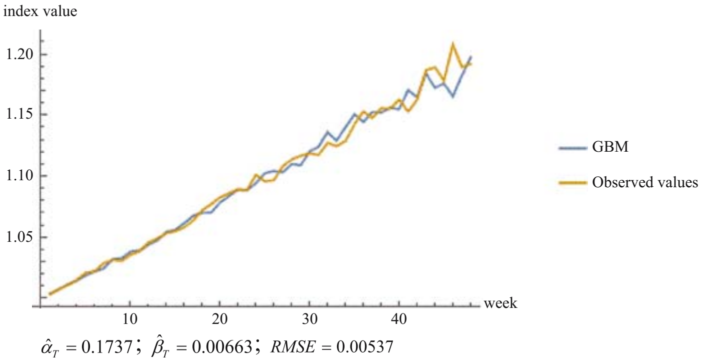


Fig. 5. Observed average normalised price process (Observed values) and sample realisation of the matched GBM price model (GBM).

results for the expected values of the sample indices obtained under the GBM model were then compared to the sample index values determined with Equations (1) and (2) based on a sample of all prices. All the results are presented in Figure 6 and Table 2. The small differences that we can observe in this comparison suggest not only that the adopted price

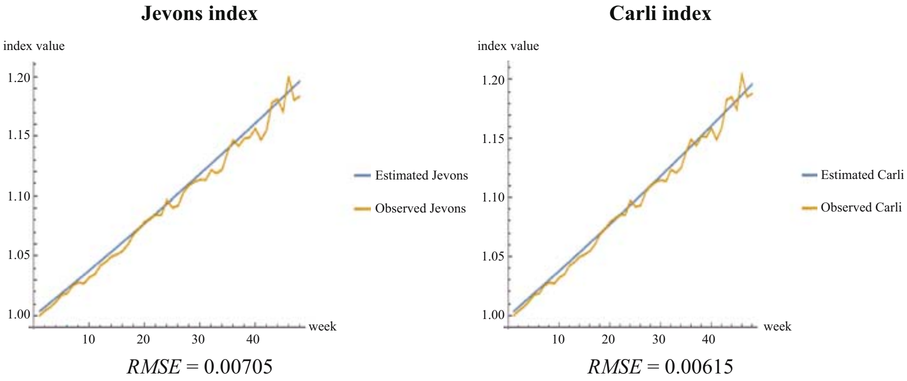


Fig. 6. Comparison of expected values of the Jevons and Carli indices with the empirical ones (the first week is the fixed base time period).

Table 2. Basic characteristics of the sample Jevons and Carli indices.

Characteristics	Jevons index	Carli index
Observed index value	1.18319	1.18791
Expected value (*)	1.18969	1.18970
Standard deviation (*)	0.00271	0.00322
Mean squared error (MSE) (*)	$7.32636 \cdot 10^{-6}$	$1.03883 \cdot 10^{-5}$

(Indices calculated for the whole time interval, that is, the last week is compared to the first week).
(*) Values obtained under the GBM price model for $\hat{\alpha}_T = 0.1737$; $\hat{\beta}_T = 0.00663$.

model is correct and its parameters have been properly determined, but also indicate the practical value of the theoretical results obtained. In practice, the estimation of parameters α and β for a sufficiently long time series of prices (for which most probably $\hat{\alpha}_T \approx \hat{\alpha}_{T+1}$ and $\hat{\beta}_T \approx \hat{\beta}_{T+1}$) would allow for making short-term accurate predictions for the values of the elementary indices discussed (e.g. for one period ahead, i.e., for $t = T + 1$). Similarly, retrospectively, it would be possible to make an accurate forecast of price changes (more precisely: to determine the expected sample indices in accordance with Equations (15) and (16)) between the time periods $m_1, m_2 \in (t_1, t_T)$ which are of interest to the researcher, but for which direct price observations have not been made. For example, in the case of the analysed homogeneous group of mountain bikes, the forecasted (expected) price change after 197 days since its distribution (i.e., for $t = 197/365 = 0.5397$) is $E(P_C) = 1.09829$, $E(P_J) = 1.09828$.

5. Simulation Study

According to considerations from Section 4, we consider here a single item with prices described by a geometric Brownian motion (8) sold in N_m places (outlets) and observed in N sample monitoring points. We intend to compare expected values, biases and MSEs of the sample Jevons and Carli indices for different sample sizes. The population of N_m prices is generated in four cases: Case 5.1 with the drift $\alpha = 0.1$ and the volatility $\beta = 0.03$ (increasing prices, low volatility), Case 5.2 with $\alpha = 0.1$ and $\beta = 0.3$ (increasing prices, high price volatility), Case 5.3 with $\alpha = -0.1$ and $\beta = 0.03$ (decreasing prices, low price volatility) and Case 5.4 with $\alpha = -0.1$ and $\beta = 0.3$ (decreasing prices, high price volatility). Our simulation procedure is as follows:

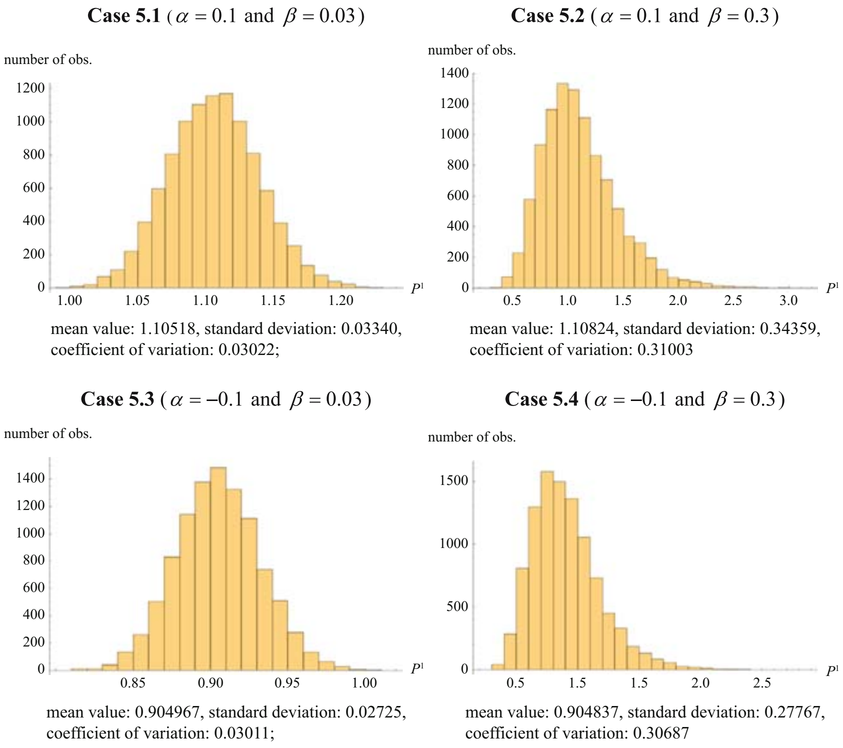
Step 1) we generate a population of $N_m = 10,000$ prices described in Equation (8) for $t = 1$ (without loss of generality, we assume that $p_i^0 = 1$ and thus the generated price p_i^1 determines the price relative p_i^1/p_i^0);

Step 2) we draw a sample of N prices $r = 1,000$ times, whereas we consider $N \in \{10, 100, 1,000\}$;

Step 3) we calculate the sample Jevons and Carli indices for each rth repetition and their distances to the theoretical expected price change $\mu = \exp(\alpha)$;

Step 4) we calculate expected values, standard deviations and coefficients of variation, as well as "empirical" biases and MSEs of the generated sample Jevons and Carli indices and we compare these results to theoretical ones obtained from the formulas presented in Subsection 3.2.

The populations of price relatives for Cases 5.1–5.4 are presented graphically in Figure 7. The sample realisations of price processes described by Cases 5.1–5.4 are presented in Figure 8. Empirical biases and MSEs of the sample Jevons and Carli indices and the analogous theoretical values are presented in Table 3.



mean value: 1.10518, standard deviation: 0.03340, coefficient of variation: 0.03022;

mean value: 1.10824, standard deviation: 0.34359, coefficient of variation: 0.31003

mean value: 0.904967, standard deviation: 0.02725, coefficient of variation: 0.03011;

mean value: 0.904837, standard deviation: 0.27767, coefficient of variation: 0.30687

Fig. 7. Histogram and XY-plot for $N_m = 10,000$ generated price relatives (population).

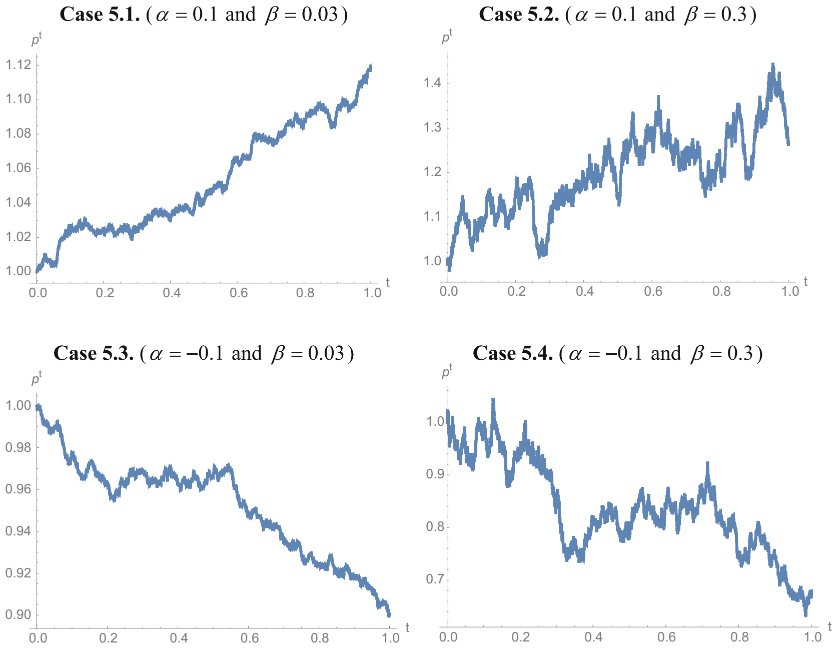


Fig. 8. Sample realisation of price processes from Cases 5.1–5.4.

Table 3. Empirical biases and MSEs of the sample Jevons and Carli indices (*) and the analogous theoretical values (**).

Case 5.1 Parameter	Sample size $N = 10$ $vol = 0.99996$		Sample size $N = 100$ $vol = 0.999996$		Sample size $N = 1000$ $vol \approx 1$	
	Theoretical value	Sample value	Theoretical value	Sample value	Theoretical value	Sample value
Bias(P_C)	0	-0.00062	0	0.00025	0	-0.00020
Bias(P_J)	-0.00004	-0.00108	-0.000004	-0.000023	-0.0000004	-0.00000069
MSE(P_C)	0.00011	0.00010	0.00001	0.00001	0.000001	0.0000001
MSE(P_J)	0.00009	0.00010	0.000009	0.00001	0.0000009	0.0000007
Case 5.2 Parameter	Sample size $N = 10$ $vol = 0.995958$		Sample size $N = 100$ $vol = 0.999555$		Sample size $N = 1000$ $vol = 0.999955$	
	Theoretical value	Sample value	Theoretical value	Sample value	Theoretical value	Sample value
Bias(P_C)	0	0.00758	0	0.00064	0	0.00041
Bias(P_J)	-0.00446	-0.03785	-0.00049	-0.0004700	-0.000049	-0.000003
MSE(P_C)	0.01150	0.01183	0.00115	0.00107	0.00011	0.00012
MSE(P_J)	0.00943	0.01213	0.000941	0.00314	0.0000941	0.0000082

Table 3. Continued.

Case 5.3 Parameter	Sample size $N = 10$ $vol = 0.99996$		Sample size $N = 100$ $vol = 0.999996$		Sample size $N = 1000$ $vol \approx 1$	
	Theoretical value	Sample value	Theoretical value	Sample value	Theoretical value	Sample value
Bias(P_C)	0	-0.00017	0	0.00006	0	-0.00013
Bias(P_J)	-0.000036	-0.00054	-0.000004	-0.00033	-0.0000004	-0.0000054
MSE(P_C)	0.00007	0.00007	0.000007	0.000007	0.0000007	0.0000007
MSE(P_J)	0.00009	0.00007	0.000009	0.000007	0.0000009	0.000001
Case 5.4 Parameter	Sample size $N = 10$ $vol = 0.995958$		Sample size $N = 100$ $vol = 0.999555$		Sample size $N = 1000$ $vol = 0.999955$	
	Theoretical value	Sample value	Theoretical value	Sample value	Theoretical value	Sample value
Bias(P_C)	0	-0.00441	0	-0.00233	0	0.00116
Bias(P_J)	-0.00365	-0.03997	-0.000403	-0.004130	-0.000040	-0.00003894
MSE(P_C)	0.00771	0.00784	0.00077	0.00073	0.00007	0.00007
MSE(P_J)	0.00943	0.00848	0.000941	0.00234	0.000094	0.000118

(*) For the fixed sample size, the given parameter is calculated for each repetition and we present a mean value of all these calculated parameter values.

(**) Theoretical values of biases and MSEs of the Jevons and Carli indices are calculated using the formulas presented in Section 4, see Equations (17), (24), and (25).

(***) The symbol 'vol' means the function $vol(t, \beta, N)$defined in (13)

6. Conclusions

There has been a general trend of replacing the Carli index with the Jevons or the Dutot formulas, and most papers recommend the Jevons index rather than the Carli index. In this article, we show some similarities and differences in the practical use of these indices.

In the **Simulation**, we generate a price population and we compare biases and mean standard errors of the sample Jevons and Carli indices under the assumption that price processes have the same drifts and volatilities. The aim of the simulation is to verify how effective approximations (24) and (29) are. In Cases 5.1 and 5.3 with low price volatility, theoretical values of biases and MSEs of indices seem to be comparable with those obtained from samples (see Table 3). In Cases 5.2 and 5.4. with high price volatility, we also obtain a satisfying estimation of biases and mean standard errors of the considered sample indices in comparison with their theoretical values (see Table 3). Approximations (22) and (29) seem to work well and we can draw some general, partly already known, conclusions. Firstly, the sample Jevons index has a negative (and as a rule small) bias. It is proven, however, that the sample Jevons index is asymptotically unbiased under the GBM price model. Secondly, the difference between MSEs of the Jevons and Carli estimates depends not only on price volatilities but also, for the fixed level of price volatilities, it may depend on the sample size (see Table 3). Our simulation confirms theoretical results that the MSE of the sample Carli index is greater than the MSE of the Jevons index in the case of a big sample size and increasing prices. This relation will change its direction if prices start to decrease. We should be aware of one more thing – although differences between expected values, variances and MSEs of the sample Jevons and Carli indices seem to be negligible, that observation concerns only one considered item. In practice, the CPI basket consists of hundreds or thousands of items (depending on the country) and the choice of the elementary formula may have serious consequences for the final CPI calculations. In the literature, the above-mentioned source of the CPI bias is called *the elementary formula bias* (White 1999). Our other simulations (which are not presented here) also confirm results obtained in Equations (15), (16), (18), and (22). Theoretical values of the expected value, as well as the standard deviation and coefficient of variation of the considered sample elementary indices, are almost identical with those obtained via the simulation study for each sample size. And finally, in the case of both indices and all the considered cases, we can observe that by increasing the size of the sample ten times, we reduce their standard deviation (and coefficient of variation) by approximately three times.

The **empirical study** (see Section 4) confirms the usefulness of the theoretical results obtained. The study from **Case 1** took into account the price of tomatoes in Poland, which is characterised not only by a high fluctuation of value in our climatic conditions, but also by seasonality. Due to the high volatility of prices, the obtained differences between the sample Jevons and Carli indices are noticeable, which once again confirms that the choice of the elementary formula of the index is extremely important in the CPI measurement. In our empirical study, the sample Carli price index has a larger variance but a smaller MSE than the sample Jevons index. As shown by our theoretical considerations and Example 1 (see Figure 1), this is not a general regularity. Interestingly, while using the Dalen's approximation of the Jevons index variance (Dalen 1999), it can be concluded that the approximate variance of the Jevons index will always be smaller than the variance of the Carli index. However, if we consider the formulas for the variance obtained in the article, see Equations (18) and (22), this relationship seems to be true, but only in the case of rising prices (i.e., when $\alpha > 0$ then $\mu_t > 1$). The empirical study presented in **Case 2** concerns prices of six models of mountain bikes that were sold in 2018 (data on transactions were downloaded from *allegro.pl*). We match the GBM price model to this data set successfully (see Figures 4–6) and, after estimating the parameters of the model for average prices, we are able to: (a) calculate expected values, standard deviations and coefficients of variation of the sample Carli and Jevons indices (see Table 2); (b) forecast future values of these indices under the assumption that the model's parameters do not change rapidly over time; (c) calculate the historical values of these indices even for time moments from which there was no information about prices. Possibilities (b) and (c) are the main advantages of using continuous time price models such as the discussed GBM model.

To sum up, although the test and economic approaches both seem to favour the Jevons index over the Carli index (the Carli index fails time reversal test and circularity), the statistical approach does not provide clear, general guidance. The author's opinion in this respect coincides with the results obtained by Levell (2015). The problem of choosing between these indices is still open – and it is troubling that, for example, under our model assumptions, the expected value of the Jevons index is very strongly dependent on the level of price volatility.

Now some final remarks: the GBM process has a number of desirable properties when it comes to modelling prices (that are currently described in Subsection 3.1). It is shown in the article, by using an empirical example, that price movements are well-described by this process and, by using simulation exercises, that the approximations of the variance, bias, and MSE of these indices are accurate. The article also shows that given this stochastic process, the difference between the Jevons and Carli indices depends specifically on the volatility of prices. Without knowledge of the exact process followed by prices, we can otherwise only say that the difference is bounded by the variance of the price relatives (a general result of the difference between arithmetic and geometric means). We also suggest that the Wiener process would be useful for forecasting and back-casting price indices (see Case 2 in the Empirical Study). However, please note that using continuous time stochastic models for real (discrete time) price cases has some limitations: (a) we must have long enough time series of high-frequency data to estimate the parameters of the used model; (b) we should verify the assumptions of the model and we have to make sure that the used model fits to the analysed data set; (c) we have to select a group of products from the CPI basket that behave according to the nature of the model. For instance, the GBM price model is not appropriate for the case of weakly seasonal goods, since the price of these goods has a rather periodical form. Thus, the assumption of the GBM model cannot be true for all the items included in CPI baskets. However, as it was shown based on examples provided, the presented methods work quite effectively in the case of prices showing a trend over time.

Please also note that there are many other interesting continuous-time stochastic models in the literature that could be applied in price modelling at the elementary level. For example, we can encounter interesting generalisations of the GBM model (Kühn and Neu 2008; You-Sheng and Cheng-Hsun 2011), a multivariate version of this model (Hu 2000), the GBM model with jumps and other extensions (Kou 2002; Hong-Bae and Tae-Jun 2015), or a more general process than the one considered in which drifts and volatilities are changeable over time. In fact, these models are much more complicated and the estimation of their parameters is not easy. In our opinion, they are more likely to be used in the nearest future, when scanner data (high frequency data) become fully available, and we know the real nature of price volatilities. The application of such extended continuous-time stochastic models for price index modelling is part of our potential future work.

7. References

Balk, B.M. 2005. "Price Indexes for Elementary Aggregates: The Sampling Approach." *Journal of Official Statistics* 21(4): 675–699. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/price-indexes-for-elementary-aggregates-the-sampling-approach.pdf (accessed September 2020).

Barlow, M.T. 2002. "A Diffusion Model for Electricity Prices." *Mathematical Finance* 12 (4): 287–298. DOI: https://doi.org/10.1111/j.1467-9965.2002.tb00125.x.

Białek, J. 2013. "Measuring Average Rate of Return of Pensions: A Discrete, Stochastic and Continuous Price Index Approaches." *International Journal of Statistics and Probability* 2(4): 56–63. DOI: https://doi.org/10.5539/ijsp.v2n4p56.

Białek, J. 2015. "Generalization of the Divisia price and quantity indices in a stochastic model with continuous time." *Communications in Statistics: Theory and Methods* 44(2): 309–328. DOI: https://doi.org/10.1080/03610926.2014.968738.

Boskin, M.S. 1996. (Chair) Advisory Commission to Study the Consumer Price Index. "Towards a More Accurate Measure of the Cost of Living." Final report for the Senate Finance Committee. Washington D.C., Available at: https://www.ssa.gov/history/reports/boskinrpt.html.

Boskin, M.S., E.R. Dulberger, R.J. Gordon, Z. Griliches, and D.W. Jorgenson. 1998. "Consumer prices in the consumer price index and the cost of living." *Journal of Economic Perspectives* 12(1): 3–26. DOI: https://doi.org/10.1257/jep. 12.1.3.

Bureau of Labor Statistics. 2001. The experimental CPI using geometric means (CPI-U-XG).

Carruthers, A.G., D.J. Sellwood, and P.W. Ward. 1980. "Recent developments in the retail price index." *The Statistician* 29(1): 1–32. DOI: https://doi.org/10.2307/2987492.

Carli, G. 1804. *Del valore e della proporzione de'metalli monetati.* In: Scrittori Classici Italiani di Economia Politica 13: 297–336. Available at: https://books.google.it/books?id=v31JAAAAMAAJ.

Cobb, B.R., R. Rumi, and A. Salmeron. 2012. "Approximation the Distribution of a Sum of Log-normal Random Variables". Paper presented at the Sixth European Workshop on Probabilistic Graphical Models, Granada, Spain, 2012. Available at: http://leo.ugr.es/pgm2012/proceedings/eproceedings/cobb_approximating.pdf.

Consumer Price Index Manual. Theory and practice. 2004. ILO/IMF/OECD/UNECE/Eurostat/The World Bank, International Labour Office (ILO), Geneva. Available at: https://www.ilo.org/wcmsp5/groups/public/--dgreports/--stat/documents/presentation/wcms_331153.pdf.

Dalén, J. 1992. “Computing Elementary Aggregates in the Swedish Consumer Price Index.” *Journal of Official Statistics* 8(2): 129–147. Available at: https://www.scb.se/-contentassets/ca21efb41fee47d293bbee5bf7be7fb3/computing-elementary-aggregates-in-the-swedish-consumer-price-index.pdf (accessed September 2020).

Dalén, J. 1994. “Sensitivity Analyses for Harmonising European Consumer Price Indices”: 147–171 in *International Conference on Price Indices: Papers and Final Report, First Meeting of the International Working Group on Price Indices*, November 1994, Ottawa, Statistics Canada. Available at: https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+1/$file/1994+1st+Meeting+-+Dal%C3%A9n+J%C3%B6rgen+-+Sensitivity+Analyses+for+Harmonising+European+Consumer+Price+Indices.pdf.

Dalén, J. 1999. “A Note on the Variance of the Sample Geometric Mean. Research Report 1.” Department of Statistics, Stockholm University, Stockholm.

Diewert, W.E. 1995. “Axiomatic and economic approaches to elementary price indexes.” Discussion Paper No. 95-01. Department of Economics, University of British Columbia, Vancouver, Canada. Available at: https://economics.ubc.ca/files/2013/06/pdf_paper_ erwin-diewert-95-01-axiomatic-economic-approaches.pdf.

Diewert, W.E. 2012. “Consumer price statistics in the UK.” Office for National Statistics, Newport. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/WS1/WS1_1_Diewert_on_Diewert_Consumer_Price_Statistics__in_the_UK_v.7__06.08__Final.pdf.

Dorfman, A.H., S. Leaver, and J. Lent. 1999. “Some observations on price index estimators.” U.S. Bureau of Labor Statistics (BLS) Statistical Policy Working Paper 29(2). Washington D.C. Available at: https://www.bls.gov/osmr/research-papers/1999/st990080.htm.

Eichhorn, W., and J. Voeller. 1976. “Theory of the Price Index.” Lecture Notes in Economics and Mathematical Systems 140. Berlin-Heidelberg-New York: Springer-Verlag. DOI: https://doi.org/10.1007/978-3-642-45492-9.

Evans, B. 2012. International comparison of the formula effect between the CPI and RPI. Office for National Statistics. Newport. Available at: https://pdfs.semanticscholar.org/6667/3228f9665c9cd31011da1ef932eee394afa9.pdf?_ga = 2.36133664.1832035111.1583498850-1342646647.1583498850. Published online in 2012.

Fenton, L.F. 1960. “The sum of log-normal probability distributions in scatter transmission systems.” *IRE Transactions on Communications Systems* 8(1): 57–67. DOI: https://doi.org/10.1109/tcom.1960.1097606.

Gajek, L., and M. Kałuszka. 2004. “On the average rate of return in a continuous time stochastic model.” Working paper. Technical University of Lodz, Poland. Available at: https://www.researchgate.net/publication/270892015_ON_THE_AVERAGE_RATE_OF_RETURN_IN_A_CONTINUOUS_TIME_STOCHASTIC_MODEL.

Greenlees, J.S. 2001. “Random errors and superlative indexes.” Working Paper 343. Bureau of Labour Statistics, Washington D.C. Available at: https://www.bls.gov/pir/-journal/gj09.pdf.

Hardy, G.H., J.E. Littlewood, and G. Polya. 1934. *Inequalities*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.2307/3605504.

Hong-Bae, K., and P. Tae-Jun. 2015. "The Behavior Comparison between Mean Reversion and Jump Diffusion of CDS Spread." *Eurasian Journal of Economics and Finance* 3(4): 8–21. DOI: https://doi.org/10.15604/ejef.2015.03.04.002.

Hu, Y. 2000. "Multi-dimensional geometric Brownian motions, Onsager-Machlup functions, and applications to mathematical finance." *Acta Mathematica Scientia* 20(3): 341–358. DOI: https://doi.org/10.1016/s0252-9602(17)30641-0.

Hull, J. 2018. *Options, Futures, and other Derivatives (10 ed.).* Boston: Pearson.

Jakubowski, J., A. Palczewski, M. Rutkowski, and L. Stettner. 2003. *Matematyka finansowa. Instrumenty pochodne.* Warszawa: Wydawnictwa Naukowo-Techniczne.

Jevons, W.S. 1865. "On the variation of prices and the value of the currency since 1782." *J. Statist. Soc. Lond.* 28: 294–320. DOI: https://doi.org/10.2307/2338419.

Kou, S.G. 2002. "A jump-diffusion model for option pricing." *Management Science* 48: 1086–1101. DOI: https://doi.org/10.1287/mnsc.48.8.1086.166.

Kühn, R., and P. Neu. 2008. "Intermittency in an interacting generalization of the geometric Brownian motion model." *Journal of Physics A: Mathematical and Theoretical* 41 (2008): 1–12. DOI: https://doi.org/10.1088/1751-8113/41/32/324015.

Levell, p. 2015. "Is the Carli index flawed?: assessing the case for the new retail price index RPIJ." *J. R. Statist. Soc. A* 178(2): 303–336. DOI: https://doi.org/10.1111/rssa.12061.

McClelland, R., and M. Reinsdorf. 1999. "Small Sample Bias in Geometric Mean and Seasoned CPI Component Indexes." Bureau of Labor Statistics Working Paper No. 324, Washington D.C. Available at: https://stats.bls.gov/osmr/research-papers/1999/pdf/ec990050.pdf.

Meade, N. 2010. "Oil prices – Brownian motion or mean reversion? A study using a one year ahead density forecast criterion." *Energy Economics* 32 (2010): 1485–1498. DOI: https://doi.org/10.1016/j.eneco.2010.07.010.

Moulton, B.R., and K.E. Smedley. 1995. "A comparison of estimators for elementary aggregates of the CPI." *Paper present at Second Meeting of the International Working Group on Price Indices.* Stockholm, Sweden. Available at: https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/8e98d9c3-d6e9363eca25727500004401/$FILE/1995%202nd%20Meeting%20-%20A%20comparison%20of%20esti.

Nwafor, C.N., and A.A. Oyedele. 2017. "Simulation and Hedging Oil Price with Geometric Brownian Motion and Single-Step Binomial Price Model." *European Journal of Business and Management* 9(9): 68–81. Available at: https://researchonline.gcu.ac.uk/files/24776117/Simulation_of_Crude_Oil_Prices_EJBM_Vol.9_No.pdf.

Office for National Statistics. 2013. National Statistician announces outcome of consultation on RPI. Office for National Statistics, Newport. Available at: https://webarchive.nationalarchives.gov.uk/20160111163943/http:/www.ons.gov.uk/ons/dcp29904_295002.pdf.

Oksendal, B. 2003. *Stochastic Differential Equations: An Introduction with Applications.* Berlin: Springer. DOI: https://doi.org/10.1007/978-3-642-14394-6_5.

Privault, N. 2012. "An Elementary Introduction to Stochastic Interest Rate Modeling." *Advanced Series on Statistical Science & Applied Probability: Volume 16, Word Scientific.* DOI: https://doi.org/10.1142/8416.

Ross, S.M. 2014. “Variations on Brownian Motion.” In *Introduction to Probability Models (11th ed.),* 612–14. Amsterdam: Elsevier.

Schultz, B. 1995. “Choice of price index formulae at the micro-aggregation level: The Canadian Empirical evidence.” (Version updated in June 1995) Paper presented at the first meeting of the *Ottawa Group on Price indices*, Canada. Available at: https://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+1/$file/1994%201st%20-Meeting%20-%20Schultz%20Bohdan%20-%20Choice%20of%20Price%20Index%20Formulae%20at%20the%20Micro-Aggregation%20Level%20The%20Canadian%20Empirical%20Evidence%202nd%20Edition.pdf.

Silver, H., and S. Heravi. 2007. “Why elementary price index number formulas differ: Evidence on price dispersion.” *Journal of Econometrics*, 140 (2007): 874–883. DOI: https://doi.org/10.1016/j.jeconom.2006.07.017.

UK Statistics Authority. 2013. Consultation on the Retail Prices Index. *Statement. UK Statistics Authority, London*. Available at: http://www.statisticauthority.gov.uk/news/statement-consultation-on-the-retail-prices-index-10012013.pdf.

Von der Lippe, p. 2007. Index Theory and Price Statistics. *Peter Lang Verlag*. DOI: https://doi.org/10.3726/978-3-653-01120-3.

White, A.G. 1999. “Measurement Biases in Consumer Price Indexes.” *International Statistical Review*, 67(3): 301–325. DOI: https://doi.org/10.1111/j.1751-5823.1999.tb00451.

Yu-Sheng H., and W. Cheng-Hsun 2011. “A Generalization of Geometric Brownian Motion with Applications.” *Communications in Statistics – Theory and Methods*, 40(12): 2081–2103. DOI: https://doi.org/10.1080/03610921003764167.

Received November 2020
Revised May 2019
Accepted March 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 763–802, http://dx.doi.org/10.2478/JOS-2020-0038

Developing Land and Structure Price Indices for Ottawa Condominium Apartments

Kate Burnett-Isaacs[1], *Ning Huang*[1], *and W. Erwin Diewert*[2]

Measuring the service flow and the stock value of condominium apartments in Canada and decomposing these values into constant quality price and quantity components is important for many purposes. In addition, the System of National Accounts requires that these service flows and stock values for condos be decomposed into constant quality land and structure components. In Canada and most other countries, such a land and structure decomposition of condominium apartment sale prices does not currently exist. In this article, we provide such a decomposition of condominium apartment sales in Ottawa for the period 1996–2009. Specific attention is paid to the roles of communal land and structure space on condominium apartment unit selling prices. Key findings include methods to allocate land and building space to a single condominium unit, identifying the characteristics that best explain condominium prices, and developing an average depreciation rate for condos for the 14-year time period.

Key words: Condominium apartment price indices; land and structure price indices; hedonic regressions; net depreciation rates; system of national accounts.

1. Introduction

Over the last 15 years, the condominium apartment sector has been a growing component of the Canadian residential property market (Read-Hobman 2015). To accurately measure the economic activity in this sector, Statistics Canada is developing a New Condominium Apartment Price Index (NCAPI) and a Resale Residential Property Price Index for condominium apartments (RRPPI Condo). The use of the NCAPI and RRPPI Condo in statistical programs, for example as a deflator in various components of Gross Domestic Product and as an input into the Consumer Price Index, requires price indices for the total (land and structure components), as well as separate price index series for land and structures. Data on separate land and structure values are difficult to come by, resulting in a knowledge gap in condominium apartment information that the NCAPI and RRPPI Condo currently cannot fill.

[1] Producer Prices Division, Statistics Canada, Ottawa, Canada, K1A 0T6. Emails: kate.burnett-isaacs@canada.ca and ning.huang2@canada.ca

[2] School of Economics, University of British Columbia, Vancouver B.C., Canada, V6T 1Z1. Email: erwin.diewert@ubc.ca.

Acknowledgments: The authors thank Heidi Ertl, Mary Beth Garneau, Gaétan Garneau and Clément Yélou at Statistics Canada and Price Measurement Advisory Committee members Jan de Haan, François Des Rosiers, Pierre Duguay, John Mallon and Mick Silver for their helpful discussions and comments. The authors also thank Amy Jiang, Simon Trottier and Ross Beck-MacNeil for their technical help on the data treatment. The views expressed in this article are solely those of the authors and do not necessarily reflect those of Statistics Canada. The last author gratefully acknowledges the financial support of the SSHRC of Canada and the Australian Research Council (LP0884095, DP150100830).

In order to decompose a condominium apartment unit price into separate land and structure components, this article uses the Builder's Model developed by Diewert and Shimizu (2016). This hedonic model suggests that the value of a condominium unit is equal to the sum of the value of its land and structure components, where each component is impacted by different characteristics and factors. As proposed by Davis and Palumbo (2008), the structure component can be viewed as the current depreciated replacement cost to build the structure. The land component measures the impact that location and neighbourhood amenities, in addition to land size, have on the total price of a condominium apartment unit. The Builder's Model used in the article starts with finding the determinants of land prices, which are estimated as a "residual" of the total value of the property minus the structure component. Using the residual as an estimate for land value is crucial for choosing the factors affecting the land prices.

This study is an application of the work first developed by Diewert and Shimizu (2016), where we tackle the challenge of allocating communal land to a single condominium unit. This article demonstrates that their basic framework can be applied in another city's data sets, in this case Ottawa, Canada. We extend their work by better accounting for the existence of communal space in the building. The challenge here is that individual condominium units share communal structural space with the rest of the building. These communal building areas include lobbies, hallways, party rooms, gyms, pools, parking lots, etc. The cost of these amenities need to be allocated to the individual property units and incorporated into the Builder's Model. In addition, we test many characteristics in our Builder's Model that are common across countries, as well as some that are unique to Ottawa. This process verifies that the characteristics chosen by Diewert and Shimizu (2016) are the most important factors affecting land and structure condo prices. This is a very useful result for other statistical agencies in that it is not necessary to collect a large number of condo characteristics in order to obtain useful overall condo price indices and useful sub-indices for the land and structure components. Lastly, we show that our hedonic regressions can generate useful estimates of structure (net) depreciation rates for wear and tear depreciation of the structure.

In this study, we focus on the Ottawa high-rise condominium apartment market from 1996 to 2009. The article is broken down in the following manner: Section 2 explains the data used in this study; Section 3 introduces the Builder's Model and how it must be adapted to incorporate land and communal building space for condominium apartments; Section 4 focuses on finding the main determinants of land prices; Section 5 introduces various structural variables to the Builder's Model; Section 6 explains the land, structure and total property index series derived from our proposed hedonic model and Section 7 concludes.

2. Data

The source of data for this study is a combination of a residential property price research data set, a City of Ottawa building characteristics data set and some internet data sources. This research data set was developed for new and resale condominium apartment units for the 55 quarters from Q1 1996 to Q3 2009 based on the data availability. The term condominium apartment is defined by Statistics Canada's National Household Survey as a private residential complex in which dwellings are owned individually, while land and common elements are held in joint ownership with others. When we define a condominium apartment in this article, it does not include single-family homes or row houses that have condominium type ownership. More specifically, we focus on high-rise condos, which are defined as those condo buildings with five or more floors. This threshold was chosen because buildings with four or less floors are built similarly to single-family houses, with higher wood content than high-rise buildings that are built with more glass and concrete materials. The data set contains unit characteristic variables such as number of bedrooms and bathrooms, the type of heating fuel, floor covering, the story that the unit is on and unit square footage; land characteristics such as location of the condominium building described by the Forward Sortation Area (FSA), land size and excess land; and building structure characteristics include the building size, building height, unit height, and the total number of units in building.

Outlier detection was conducted for the main variables unit living area, selling price, bedroom, bathrooms, and age due to misreporting and unique units. To effectively detect outliers, we examined the histograms, summary statistics and detailed percentiles of living area, bedrooms and bathrooms. A simple cut-off technique was applied to remove the extreme values. For instance, the units with either top or bottom 1% living areas are removed from the sample; and the maximum values for bedrooms and bathrooms are trimmed off. The age restriction of 50 years was chosen because buildings older than this age will most likely have gone through a major renovation. Since we use age of the building to estimate depreciation, including buildings with major renovations would not provide accurate results. The selling prices were trimmed off unsymmetrically due to the fact that the distribution of selling prices are positive skewed. Different cut-off values were tested to determine how sensitive the final indices were to the choice of cut-off values with the help of pooled time dummy hedonic model. The final data set includes observations with the following characteristics:

- Living area between 300 and 1500 square feet (sqft),
- Selling price between bottom 1% and top 5% by year of sale,
- One to four bedrooms,
- One to three bathrooms, and
- Age $<$ 50 years.

Descriptive statistics for sales price by year is outlined in Table 1 and the main characteristic variables that will be used in our analysis are listed in Table 2. It can be seen that, even after range trimming, there is still a great deal of variation in the explanatory variables such as selling prices, total residential building area and the lot size of the condo buildings.

3. The Builder’s Model for Condominium Apartments

The Builder’s Model is based on the expected cost of building a property, either a single family home or a condominium apartment unit. This model suggests that the selling price of a property that has a newly built structure on it is driven by the cost of producing said property. Thus, the hedonic form of the Builder’s Model states that property price is equal

Table 1. Descriptive statistics for selling price by year.

Year	Frequency	Mean	Standard deviation	Minimum	Maximum	Median	Skewness
1996	468	95,909.62	30,847.31	40,000	185,000	89,000	0.8747556
1997	565	94,335.69	30,011.55	36,000	191,000	90,000	0.7427555
1998	541	89,813.8	29,232.25	32,000	173,000	87,000	0.5283304
1999	664	95,981.17	30,968.67	38,000	195,000	90,000	0.852223
2000	728	104,125.2	31,876.98	45,000	210,000	101,250	0.718588
2001	734	126,339.3	37,253.06	61,000	239,000	123,000	0.6708337
2002	775	151,408.6	44,495.66	74,900	291,000	147,500	0.6225587
2003	659	169,487.9	45,990.41	94,000	300,000	163,500	0.7674735
2004	730	184,568.9	50,428.61	104,000	344,900	174,925	0.8914522
2005	754	190,905.7	52,218.65	107,000	360,000	180,000	0.9160729
2006	878	199,025.4	62,704.01	109,900	398,000	182,500	1.081514
2007	967	211,261.4	68,310.64	100,000	430,000	193,000	1.053023
2008	808	230,794.8	74,441.29	113,000	455,000	215,000	0.9118591
2009	735	243,615.1	77,570.45	118,000	460,000	230,000	0.662988

Table 2. Descriptive statistics for key characteristic variables.

Variable	Frequency	Mean	Standard deviation	Minimum	Maximum	Mode
Unit living area (sqft)	10,006	667.57	155.65	300.06	1,495.64	
Lot size (sqft)	10,006	92,180.86	65,209.38	2,029.00	26,8021.13	
Building size (sqft)	10,006	222,685.20	108,551.00	15,021.00	614,823.18	
Age (years)	10,006	20.63	8.98	0	42	23
Height of building (stories)	10,006	16.32	6.74	5	32	11 and 12
Story of unit	10,006	8.44	5.88	1	28	3
Bedrooms (number)	10,006	1.92	0.51	1	4	2
Bathrooms (number)	10,006	1.52	0.51	1	3	2

to the quality-adjusted cost of land per square foot (α_t) times the square footage of land (TL_{tn}) plus the quality adjusted structure cost per square foot (β_t) times the square footage of the structure (S_{tn}) for n = 1,. . .N_t, where N is the total number of observations, for a given time period (t). The Builder's Model can be approximated by the following hedonic regression model, estimated using maximum likelihood estimation, with an error term (ε_{tn}) that is assumed to be normally distributed with a mean of zero and a constant variance:

$$P_{tn} = \alpha_t TL_{tn} + \beta_t S_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{1}$$

The above model applies to new properties. A constant term is not included because of the assumption that if there is no land and no structure, the property has no value. To incorporate depreciation that occurs in older structures, which devalues the structure in the absence of renovations, the Builder's Model can use information on the age of the structure (A_{tn}) in order to estimate a net geometric depreciation rate (δ_t) as the structure ages one period using the following model:

$$P_{tn} = \alpha_t TL_{tn} + \beta_t (1 - \delta_t)^{A_{tn}} S_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{2}$$

In trying to estimate Equation 2, as discussed in Handbook on Residential Property Price Indexes – RPPI Handbook (Diewert 2013), multicollinearity between the land and structure variables warrants the use of a construction cost index to proxy for the change in cost of building the structure. Multicollinearity occurs when two or more independent variables are correlated with each other. This can cause estimates to be unstable and difficult to interpret, with potentially incorrect signs or magnitudes. In this study, the price per square foot of a condominium is set equal to that from Statistics Canada's Apartment Building Construction Price Index (ABCPI). Using the ABCPI will allow our model to be consistent with the Canadian System of Macroeconomic Accounts estimates for the value of new construction. The use of this variable is based on the assumption that the movement of condominium apartment building costs approximates those for non-condominium apartment buildings. This assumption is based on the grounds that increasingly, rental apartment buildings are constructed with similar finishes as condos. This price per square foot is then indexed using the ABCPI for period t to get an estimated cost per square foot of structure space (PS_t) for each quarter from Q1 1996 to Q3 2009. The resulting hedonic model is:

$$P_{tn} = \alpha_t TL_{tn} + \beta PS_t (1 - \delta_t)^{A_{tn}} S_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{3}$$

where β coefficient now represents a general quality adjustment factor to the structure area and is constant over time. To apply the Builder's Model to condominium apartment units, we need to make additional considerations that would not be found in the Model for, say, single-family homes. The main considerations are how to address the roles of communal land and structure space on the selling price of a condominium apartment unit.

3.1. Allocating the Unit's Land Share: Method 1

In our data set, the variable for unit area is used to estimate the structure component for the unit only. However, land size is given for the whole building and not the single unit.

Therefore, land size must be allocated appropriately to a single condo unit. The preliminary assumption is that each unit in the building equally enjoys the whole land area, therefore the land should be divided equally by all units in the building (TU_{tn}):

$$L_{tn} = \left(\frac{1}{TU_n}\right) TL_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{4}$$

In Section 4, we explore the alternative land imputation methods of allocated land proportional to the size of the unit and a weighted average of the equal distribution and size distribution methods.

3.2. Allocating the Unit's Share of Communal Space

A unique feature of condominium apartments, versus other building structures such as single-family homes, is that a single unit shares communal space with other units in the same building. This space and the amenities in it are accounted for, in addition to such sources of revenue as condo fees, in the selling price of the unit. However, when using the Builder's Model, the floor space of the condominium unit only covers the area of the privately owned space.

Explicit values of communal space are difficult to obtain and are often not reported in databases such as listing services, land registries or property assessment. Therefore, we need to estimate the proportion of the building space that is communal by other means. Consultations were conducted with the construction industry and an estimate for communal space was calculated using the apartment building specifications from the 2004 model used in the ABCPI. This model is representative of a building that can be built anywhere in Canada. From these two sources, it was determined that about 20–30% of the total floor space of the building is allocated to communal areas. We tested the sensitivity of this assumption and found that there was very little difference in the estimates of the Builder's Model using 20, 25 or 30% values for communal space. Appendix 1 (Subsection 8.1) illustrates the result of using 20, 25 and 30% communal space on model 22. Therefore, the hedonic models that follow will use an estimate for communal space of 25% of the total building area.

The floor area of the unit represents privately owned structure space. To capture all of the structural space allocated to a condominium unit, including communal space, the privately owned space in our model must be blown up by a factor that represents communal space. Since we are using the estimate of 25% of the building as communal space, 75% of the building is private space and so can be estimated by the unit floor space. To include that extra 25% of communal space in our model, structural space is estimated by

$$(1/0.75)S_m \text{ or } (1.33)S_m.$$

The amenities contained in this 25% communal space can differ between buildings and could be a factor affecting the price of a condominium unit. Tests were conducted to determine the impact that indoor and outdoor parking, fitness facilities, party rooms and indoor pools had on the price of a condominium unit. Since these communal amenities were shown to have a marginal impact on property values using the Builder's Model, and the results of regressions using communal amenities are not included in this study. In this study, a small impact is defined as improving the regression's Log Likelihood value by a small amount relative to the increase in the number of parameters.

3.3. A Preliminary Builder's Model

In order to get initial land value estimates, which will be discussed in Section 4, the depreciation rate was set equal to 2%. This is the estimate used by the Canadian System of Macroeconomic Accounts and productivity analysis at Statistics Canada for all residential housing depreciation rates. Including the unit's share of land (L_{tn} defined by Equation (4)), the communal space blow-up factor (1.33), the price per square foot of structural floor space (PS_t) and the exogenous annual depreciation rate (0.02), the model defined by Equation (2) becomes the following model:

$$P_{tn} = \alpha_t TL_{tn} + (1.33)\beta PS_t(1-0.02)^{A_{tn}} S_{tn} + \varepsilon_{tn}; \ \ \mathrm{t} = 1, \ldots, 55; \ \ \mathrm{n} = 1, \ldots, N_t. \quad (5)$$

The R^2 value for this model turned out to be 0.6774, which indicated that there was room for improvement in the model. However, the results for this model were not plausible. Most αt values, which are the estimates for average price of land per square foot, were negative. Negative prices cannot exist in this context. Also, the β coefficient estimate, representing a general structure quality adjustment factor, was 5.02528 (t-stat = 330.25). Our assumption is that the model can account for almost all quality adjustment to the structure implying that β should be close to a value of 1. Thus, the very large estimate for β has led to α_t estimates which are too small to be credible, as demonstrated in Appendix 2 (Subsection 8.2).

4. The Determinants of Condominium Land Prices

To improve the results of the model defined by Equation (5), we set β equal to 1 which allowed us to focus on finding the main determinants of land prices. To do this, we temporarily took imputed land value to be the dependent variable of our hedonic model. We derive an estimate for imputed land value for property n in period t, (LV_{tn}), by subtracting our imputed structure value (SV_{tn}) from the total property-selling price (P_{tn}):

$$LV_{tn} = P_{tn} - SV_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \quad (6)$$

Our imputed structure value is approximated by:

$$SV_{tn} = (1.33)PS_t(1-0.02)^{A_{tn}} S_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \quad (7)$$

The above estimates for land value were then used as the dependent variable for the models that are described in this section. The baseline model we will use to begin our analysis is that land value in period t can be modeled by the price of land per square foot (α_t) multiplied by the equally distributed land area defined by Equation (4) above, (L_{tn}):

$$LV_{tn} = \alpha t L_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \quad (8)$$

This model gave us a starting point to assess the impact of additional land characteristics on the goodness of fit of the subsequent models. The very simple model defined by Model (8) had an R square value of (0.673 and a log likelihood value (LL) of (127,824).

This negative R square value is a result of not having a constant term in Model (8). Given the nonlinear nature of this and subsequent models, goodness of fit will be determined by the combined improvement in the LL and the R-square values.

4.1. Introducing Postal Code Dummy Variables

The results of Model (8) clearly suggest that there needed to be an improvement in how we modeled land prices. The price of any property is heavily impacted by location. To capture this relationship, we use Forward Sortation Area dummy variables ($FSA_{tn,i}$) in our hedonic model. A forward sortation area (FSA) is a geographic unit based on the first three characters in a Canadian postal code. These 22 dummy variables are defined as:

$$\begin{aligned} FSA_{tn,i} &= 1 \text{ if observation n in period t is in Forward Sortation Area i;} \\ &= 0 \ \text{ otherwise.} \end{aligned} \quad . \qquad (9)$$

By adding the Forward Sortation Area dummy variables to Model (8) we can account for how the land prices change based on the location of the condominium:

$$LV_{tn} = \alpha_t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})L_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \qquad (10)$$

where the land size by unit (L_{tn}) is defined as in Equation (4). The 55 α_t parameters and 22 θ_i parameters cannot all be identified. Therefore, we normalized $\alpha_1 = 1$. With $\alpha_1 = 1$, the value of all other α_t estimates represent the percentage change in land value due to the change in time from period 1 to period *t*. This is the definition of a price index and so we can use the parameter estimates of land price to create our land price index. The R-square for this model was 0.0984 and the LL was (124,732, which was a large 3,092 improvement from Model (8), validating our assumption that location has a significant impact on the land prices in Ottawa.

4.2. Alternative Land Value Distribution Methods

Up to this point, our models assumed that land was equally distributed to each condominium apartment unit. However, land could also be allocated to a single unit proportionally to the floor size of the unit or to a combination of equal and proportional allocation.

Land can be allocated to a single unit proportionally to its size compared to the rest of the building. Like in the case of condo fees, where larger units pay higher fees and thus contribute more to funding communal spaces, the logic in this assumption is that the larger units should have a larger share of the land. Proportional land size (L_{tn}) is defined as:

$$\overset{*}{L}_{tn} = \left(\frac{S_{tn}}{TS_{tn}}\right) TL_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \qquad (11)$$

Replacing the L_{tn} variable in equations (10) by the proportional land variable L^{*}_{tn} defined in (11), our model becomes:

$$LV_{tn} = \alpha t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})\left(\frac{S_{tn}}{TS_{tn}}\right)TL_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{12}$$

However, the R-square and LL values from this model, −0.0956 and −125,172 respectively, were worse than those of Model (10).

Given that Model (12) provides worse results than Model (10), we need to find a different method to determine the land share of a single unit. An alternative method is to distribute the total land among the units in the building by a *weighted average* of the equal and the proportional allocation:

$$L^{**}_{tn} = \left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right]TL_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{13}$$

The ρ coefficient is estimated in Model (14) below:

$$LV_{tn} = \alpha_t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right]TL_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \tag{14}$$

The estimate of $\hat{\rho}$ was 0.24021 (t stat = 11.42), therefore placing a higher weight towards equally distributing the land to a single unit. This makes sense given the poor performance of proportionally distributing land alone, as was found in Model (12).

The R-square of Model (14) was 0.1025 and the LL was −124,178, which is an improvement of 994 on Model (12) and an improvement of 554 on Model (10). Therefore, subsequent models of land value and total property price will use this weighted land distribution method. The intuition for including the weighted average is that the more units in a building, the greater the heterogeneity in size of the units within the building, and so the greater the impact a unit of a larger size, or larger proportional size, has on price.

4.3. Introducing the Height of the Unit

Our expectation is that a unit on a higher floor will have a better view than those on lower floors. The view can be thought as the vertical dimension of land. Therefore, the floor the unit is on, or the height of the unit (H_{tn}), impacts the price of land for a condominium unit. The variable H_{tn} was added to Model (14) as a continuous variable because it represented the response of the unit's price to a change in height of the unit in a more parsimonious way than using height dummy variables. Thus, we added $(1 + \gamma(H_{tn} - 1))$ to Model (14) and obtain Model (15):

$$LV_{tn} = \alpha_t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + \varepsilon_{tn}; \tag{15}$$

$$t = 1, \ldots, 55; \quad n = 1, \ldots, N_t.$$

Even though H_{tn} is a continuous variable, we still normalize the impact that the height of the unit has over the lowest floor observed in our data, in this case the first story. The predicted value of land price will not be affected by those observations corresponding to a unit sold on the first floor. For any unit on a floor above the first floor, the land price will increase by γ for each story. Our estimate for $\hat{\gamma}$ was 0.04121 (t stat $= 29.22$). Therefore, the predicted land price of a condominium unit increased by 4.12% for every story above the first floor. The R-square of this model is 0.1893 and the LL was −123,671, an increase of 507 over Model (14).

4.4. Introducing the Number of Units in the Building

In order to build a condominium apartment building, land needs to be zoned for the type and size of building. A building with more units will cost more in zoning fees and builders will pass these extra costs on to consumers. To test the extent to which an extra unit impacts the sale price of a condominium unit, we introduce the total number of units (TU_{tn}) into Model (15) in a similar fashion to the height of the unit (H_{tn}) as a continuous variable. Again, we normalized the impact that an extra unit will have above the minimum number of units found in a building in our data set. In this case, that minimum number of units is nine. We updated Model (15) with multiplicative factor $(1 + \omega(Tu_{tn} - 9))$ where ω represents the percentage change in land value due to an increase of one extra unit in the total number of units found in a building. If the building has nine units in it, its land value will be unaffected.

Our new hedonic model is as follows:

$$LV_{tn} = \alpha_t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))$$

$$\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + \varepsilon_{tn}; \tag{16}$$

$$t = 1, \ldots, 55; \quad n = 1, \ldots, N_t.$$

The R-square value of Model (16) was 0.2944 and the LL was −122,980, which is an increase of 691 over Model (15). The resulting estimate for $\hat{\omega}$ is 0.009432 (t stat $= 34.22$) indicating that one extra unit in the building will increase the value of land for a single condominium unit by 0.94%.

4.5. Introducing the Height of the Building

Certain neighbourhoods are zoned for tall buildings, such as downtown areas. These buildings are generally more expensive, but to what extent is that because these buildings are tall or because they are in downtown? In Model (10) we accounted for location, so now we want to determine the impact building height has on land values. To measure this effect, we introduce four height of building dummy variables to Model (16) based on the quartiles of the total building height (TH_{tn}) found in our data set. Group 1 is defined as containing observations for $TH_{tn} < 11$ stories; group 2 contains observations where $11 \leq TH_{tn} < 15$; group 3 contains observations where $15 \leq TH_{tn} < 22$ and group 4 contains observations where $TH_{tn} \geq 22$. The quartile groupings were chosen to ensure that there were enough observations for all dummy variables. The total building height dummy variable is defined as:

$$\begin{aligned} TH_{tn,j} &= 1 \text{ if observation n in period t is in total building height group j;} \\ &= 0 \ \text{ otherwise.} \end{aligned} \quad . \qquad (17)$$

It is important to note that the height of the building does not change over time. Since our observations are observed for a given time t, we include the time subscript in our variable definition. The hedonic model including the total height dummy variables is as follows:

$$\begin{aligned} LV_{tn} = {} & \alpha_t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))(\textstyle\sum_{j=1}^{4} \vartheta_j TH_{tn,j}) \\ & \left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + \varepsilon_{tn}; \qquad (18) \\ & \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t. \end{aligned}$$

The four total building height parameters (ϑ_j), the 22 Forward Sortation Area dummy parameters (θ_i) and the 55 land price parameters in Model (18) cannot be all identified, therefore we apply the following normalizations on these parameters:

$$\alpha_1 = 1; \ \vartheta_1 = 1. \qquad (19)$$

The R-square value for Model (18) was 0.3594 and the LL was –122,498, an increase of 482 over the LL of Model (16). The estimated total building height parameters increase as the building height increases, suggesting that even accounting for location, building height increases land prices. Introducing height of the building after height of the unit also brings in the concept of relative height. A unit on the 10th floor of an 11-story building will have a different price than a unit on the 10th floor of a 20-story building because of the different view that it would offer after controlling the location of the building. Taller buildings may add benefit to the view because they are taller than surrounding buildings or match the height of surrounding buildings.

4.6. Introducing Excess Land

The excess land surrounding a condominium building can incorporate many land characteristics that we cannot account for given our data. Excess land is measured as the total land plot area minus the building footprint (total building floor area divided by number of floors in the building). If a building has a large amount of excess land it could mean this excess property contains amenities such as outdoor parking, outdoor pools, parks and pathways. We do not have these amenity characteristic variables in our data set and so excess land can account for some of these extra land features. We created four excess land dummy variables ($EL_{tn,m}$) based on the quartiles of excess land size found in our data: group 1 is made up of observations where $EL_{tn} < 22{,}254$ square feet; group 2 contains observations where $22{,}254 \leq EL_{tn} < 76{,}424$ square feet; group 3 contains observation where $76{,}424 \leq EL_{tn} < 124{,}269$ and group 4 contains observations where $EL_{tn} \geq 124{,}269$. The quartile ranges were chosen to ensure that there were enough observations for each grouping of excess land. The excess land dummy variables are created as follows:

$$\begin{aligned} EL_{tn,m} &= 1 \text{ if observation n in period t is in excess land group m;} \\ &= 0 \text{ otherwise;} \end{aligned} \quad . \qquad (20)$$

In addition to the normalizations imposed for Model (18), we set $\sigma_1 = 1$, so that all remaining parameters can be identified. We added these $EL_{tn,m}$ dummy variables to Model (18) to get the following model:

$$\begin{aligned} LV_{tn} = \alpha t\left(\sum\nolimits_{i=1}^{22} \theta_i FSA_{tn,i}\right)(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))\left(\sum\nolimits_{j=1}^{4} \vartheta_j TH_{tn,j}\right) \\ \left(\sum\nolimits_{m=1}^{4} \sigma_m EL_{tn,m}\right)\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + \varepsilon_{tn}; \end{aligned} \qquad (21)$$

$$t = 1, \ldots, 55; \quad n = 1, \ldots, N_t.$$

where $\alpha_1 = 1$; $\vartheta_1 = 1$; and $\sigma_1 = 1$.

The R-square value for Model (21) was 0.6199 and the LL was (119,898, which is a 2,600 improvement over the LL of Model (18). Even though excess land has a significant impact, the results are not what we originally expected. Due to the amenities and potential view that more excess land could offer a condominium unit, we assumed that more excess land would increase the unit price of land. However, as one can see in Table 3, although the excess land is positively related to the price of land, the estimated $\hat{\sigma}_m$ decreased as the excess land gets bigger, which seems counterintuitive to our assumption. The same result was found in Diewert and Shimizu (2016). The significant increase in LL with the inclusion of excess land signifies that the presence of extra land is an important factor in determining the sales price of a condominium apartment. However, the decrease in the estimated $\hat{\sigma}_m$ suggests there might be decreasing returns to scale for excess land. Moreover, with further examination of the data, it could be seen that, generally speaking, condominium buildings with relatively small excess land are concentrated in the downtown core of Ottawa; while buildings with larger excess land are mainly located far away from the central area.

4.7. Estimates for the Determinants of Land Value

Table 3 displays the estimated coefficients and T statistics for key determents of land value from Model (14) to (21), where $\hat{\gamma}$ is the estimate for change in land value of a unit due to an increase in the floor that unit is on, $\hat{\omega}$ is the estimate for the change in land value for a unit

Table 3. Estimates (and t statistics) of key determinants of land value from Models 14 to 21.

Determinants of land value	Coefficient	Model 14	Model 15	Model 16	Model 18	Model 21
Land distribution weight	$\hat{\rho}$	0.240214 (11.42)	0.232094 (11.52)	0.279215 (16.01)	0.41663 (27.23)	0.51334 (63.49)
Unit height	$\hat{\gamma}$	-	0.041214 (29.22)	0.019099 (18.16)	0.010467 (10.75)	0.00821 (13.71)
Number of units in the building	$\hat{\omega}$	-	-	0.009432 (34.22)	0.003473 (19.43)	0.00958 (40.04)
11 Height of building $<$ 15 stories	$\hat{\vartheta}_2$	-	-	-	1.0005 (74.62)	1.1668 (109.17)
15 Height of building $<$ 22 stories	$\hat{\vartheta}_3$	-	-	-	1.13318 (55.86)	1.45254 (80.2)
Height of building 22 stories	$\hat{\vartheta}_4$	-	-	-	1.71641 (45.97)	1.71222 (72.32)
22,254 Excess land $<$ 76,424 square feet	$\hat{\sigma}_2$	-	-	-	-	0.55125 (131.43)
76,424 Excess land $<$ 124,269 square feet	$\hat{\sigma}_3$	-	-	-	-	0.27964 (86.2)
Excess land 124,269 square feet	$\hat{\sigma}_4$	-	-	-	-	0.18943 (59.74)
R square		0.1025	0.1893	0.2944	0.3594	0.6199
Log likelihood		-124,178	-123,671	-122,980	-122,498	-119,898

due to an extra unit in the building, $\hat{\vartheta}_j$ are the parameter estimates for the total building height dummy variables, $\hat{\sigma}_m$ are the parameter estimates for the excess land dummy variables and lastly, $\hat{\rho}$ is the land distribution weight estimate. Appendix 3 (Subsection 8.3) displays the final estimates for Model (21), where $\hat{\theta}_i$ are the parameter estimates for the Forward Sortation Area dummy variables, and $\hat{\alpha}_t$ are the parameter estimates for land value for a condominium unit in period t.

With the final determinants of land value all included in Model (21), our parameters previously estimated have changed. The main parameter change to note is the land imputation weight ($\hat{\rho}$) which grew from 0.240 in Model (14) to 0.513 in Model (21). This means that in Model (21) the proportional to size land imputation method gets about the same weight as the equal share land imputation method. We also see that height of the unit ($\hat{\gamma}$) now has a lesser impact on land value at 0.82% per additional floor. Certain trends continue in Model (21). The parameter estimates for building height ($\hat{\vartheta}_j$) increased as the height increases. We also still observe that the excess land coefficient estimates ($\hat{\sigma}_m$) decreased with the size of excess land. Though this variable significantly improves our model, as was suggested by the 2,600 increase in LL with the inclusion of this variable, the resulting estimates suggest that extra land might not be a priority for consumers looking to purchase condominium apartment units.

5. Quality Adjustment Variables for the Structure Component of Condo Value

Now that we have determined the main characteristics that contribute to land prices, we can use these variables in a Builder's Model that includes both land and structure components. Thus in the model defined by Equation (22), the net geometric depreciation rate (δ) is estimated (instead of being set equal to 2%) and where the dependent variable is now the condominium property price (instead of the imputed land value):

$$P_{tn} = \alpha t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))(\textstyle\sum_{j=1}^{4} \vartheta_j TH_{tn,j})$$

$$(\textstyle\sum_{m=1}^{4} \sigma_m EL_{tn,m})\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1-\rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + (1.33)PS_t \qquad (22)$$

$$(1-\delta)^{A_{tn}} S_{tn} + \varepsilon_{tn}; \quad \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t.$$

The R-square value of Model (22) was 0.6975 and the LL was −119,869. The estimate for the depreciation rate ($\hat{\delta}$) was 0.010636 (t stat = 10.41), which is much lower than expected. Therefore, we need to consider structural quality adjustment factors, such as the number of bedrooms and the number of bathrooms, in our Builder's Model before we adopt this lower net depreciation rate.

5.1. *Introducing the Number of Bedrooms*

Even after accounting for unit size, the number of bedrooms in a condominium unit can impact its selling price. The condominium units found in our data have between one to four bedrooms. We group our observations based the number of bedrooms found in the unit: group 1 contains observations with one bedroom; group 2 observations have two bedrooms and group 3 observations have three or four bedrooms. Three and four bedrooms were grouped together due to the small sample size of units with four bedrooms. We introduce a bedroom dummy variables ($BD_{tn,k}$) into Model (22) based on the following definitions:

$$\begin{aligned} BD_{tn,k} &= 1 \text{ if observation n in period t is in bedroom group k;} \\ &= 0 \text{ otherwise.} \end{aligned} \quad . \tag{23}$$

The hedonic model accounting for the impact of bedrooms on selling price is as follows:

$$\begin{aligned} P_{tn} = \alpha t(&\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1+\gamma(H_{tn}-1))(1+\omega(TU_{tn}-9))(\sum_{j=1}^{4} \vartheta_j TH_{tn,j}) \\ &(\textstyle\sum_{m=1}^{4} \sigma_m EL_{tn,m})\left[\rho\left(\dfrac{S_{tn}}{TS_{tn}}\right)+(1-\rho)\left(\dfrac{1}{TU_{tn}}\right)\right]TL_{tn}+(1.33)PS_t \\ &(1-\delta)^{A_{tn}}(\textstyle\sum_{k=1}^{3} \tau_k BD_{tn,k})S_{tn}+\varepsilon_{tn}; \quad \text{t} = 1, \ldots, 55; \quad \text{n} = 1, \ldots, N_t. \end{aligned} \tag{24}$$

We apply the following normalization parameters to the model defined by Equation (24):

$$\alpha_1 = 1; \ \vartheta_1 = 1; \ \sigma_1 = 1; \text{ and } \tau_1 = 1. \tag{25}$$

The R-square value for Model (24) was 0.7764 and the LL was –118,364, which is a 1,505 increase in the LL from Model (22). This large increase in LL indicates that the number of bedrooms significantly impacts the selling price of a condominium unit. The coefficient estimates increased with the number of bedrooms in the unit, signifying that more bedrooms increased the sale price of a condominium apartment unit.

5.2. Introducing the Number of Bathrooms

Bathrooms are key features in any residential property and condominium apartment units are no exception. To test the exact impact that bathrooms have on the selling price of condominium units, we introduce a number of bathroom dummy variables as structure quality adjustment variables. The condominium units found in our data have between one and three bathrooms. We group our observations based on the number of bathrooms found in the condominium unit: group 1 observations have one bathroom; group 2 observations have two bathrooms, and group 3 observations have three bathrooms. We introduce bathroom dummy variables ($BT_{tn,c}$) into Model (24) based on the following definitions:

$$\begin{aligned} BT_{tn,c} &= 1 \text{ if observation n in period t is in bathroom group c;} \\ &= 0 \text{ otherwise.} \end{aligned} \quad . \tag{26}$$

The hedonic model including the number of bathrooms is as follows:

$$\begin{aligned}P_{tn} = {} & \alpha t(\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))(\textstyle\sum_{j=1}^{4} \vartheta_j TH_{tn,j}) \\ & (\textstyle\sum_{m=1}^{4} \sigma_m EL_{tn,m})\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right) + (1 - \rho)\left(\frac{1}{TU_{tn}}\right)\right] TL_{tn} + (1.33)PS_t \\ & (1 - \delta)^{A_{tn}}(\textstyle\sum_{k=1}^{3} \tau_k BD_{tn,k})(\textstyle\sum_{c=1}^{3} \varphi_c BT_{tn,c})S_{tn} + \varepsilon_{tn}; \\ & \mathrm{t} = 1, \ldots, 55; \quad \mathrm{n} = 1, \ldots, N_t.\end{aligned} \tag{27}$$

As in previous models, we need to apply normalization parameters to Model (27), in order to identify the remaining parameters. The normalizations are as follows:

$$\alpha_1 = 1;\ \vartheta_1 = 1;\ \sigma_1 = 1;\ \tau_1 = 1;\ \text{and}\ \varphi_1 = 1. \tag{28}$$

The model defined by (27) had an R-square value of 0.7848 and a LL of (118,172, which is an increase in LL of 192 from Model (25). The positive and increasing values of $\hat{\varphi}_2$ and $\hat{\varphi}_3$, at 1.328436 (t-stat = 77.88) and 1.478348 (t-stat = 22.56) respectively, suggest that the more bathrooms found in a condominium unit, the higher it will sell for.

5.3. Additional Structure Characteristics As Explanatory Variables

Other structural characteristic variables were added to the model defined by Model (27) such as hardwood floors in the unit, natural gas in the unit, the number of appliances in the unit, on-suite bathrooms, dens, balconies, whether or not the unit was a new build and condo fees. Only balconies and the presence of natural gas in the unit significantly improved the model. Significant improvement of the model in this case refers to improving the model by at least 100 LL points. Though this is not a critical value when conducting a Log Likelihood test, this threshold was chosen because adding more variables to our hedonic non-linear model made it more difficult for the model to converge. Therefore, the threshold of 100 was chosen to balance convergence with variable choice. Both the balcony and natural gas variables are grouped into two categories: group 1 are for those observations that have the structural characteristic and group 2 are those observations that do not have the structural characteristic in question. To include balconies and natural gas into our Builder's Model we introduce a balcony ($BC_{tn,y}$) and a natural gas ($NG_{tn,z}$) dummy variable to Model (27) defined as:

$$\begin{aligned}BC_{tn,y} &= 1 \text{ if observation n in period t is in group y;} \\ &= 0 \text{ otherwise.}\end{aligned} \tag{29}$$

$$\begin{aligned}NG_{tn,z} &= 1 \text{ if observation n in period t is in group z;} \\ &= 0 \text{ otherwise.}\end{aligned} \tag{30}$$

Both variables were added individually and combined to Model (27). Adding balconies only to Model (27) resulted in an R-square value of 0.7972 and a LL −117,876, which is an improvement of 296 over the LL of Model (27). Adding natural gas to Model (27) resulted in an R-square value of 0.7961 and a LL of (117,903, an improvement of 269. Since both structural characteristics individually improved the model, we introduced balcony and natural gas dummy variables to Model (27) as shown in Model (31):

$$\begin{aligned} P_{tn} = {} & \alpha t(\textstyle\sum_{i=1}^{22}\theta_i FSA_{tn,i})(1+\gamma(H_{tn}-1))(1+\omega(TU_{tn}-9))(\textstyle\sum_{j=1}^{4}\vartheta_j TH_{tn,j}) \\ & (\textstyle\sum_{m=1}^{4}\sigma_m EL_{tn,m})\left[\rho\left(\frac{S_{tn}}{TS_{tn}}\right)+(1-\rho)\left(\frac{1}{TU_{tn}}\right)\right]TL_{tn}+(1.33)PS_t \\ & (1-\delta)^{A_{tn}}(\textstyle\sum_{k=1}^{3}\tau_k BD_{tn,k}) \\ & (\textstyle\sum_{c=1}^{3}\varphi_c BT_{tn,c})(\textstyle\sum_{y=1}^{2}\pi_y BC_{tn,y})(\textstyle\sum_{z=1}^{2}\eta_z NG_{tn,z})S_{tn}+\varepsilon_{tn}; \\ & \mathrm{t}=1,\ldots,55;\quad \mathrm{n}=1,\ldots,N_t. \end{aligned} \tag{31}$$

As in previous models, every parameter cannot be identified, so we applied the following normalization restrictions to Model (31):

$$\alpha_1 = 1;\ \vartheta_1 = 1;\ \sigma_1 = 1;\ \tau_1 = 1;\ \varphi_1 = 1;\ \pi_1 = 1;\ \text{and } \eta_1 = 1. \tag{32}$$

Model (31) had an R-square value of 0.8052 and a LL $-117{,}677$, which is an improvement in LL over Model (27) by 495. The coefficient estimates for the balcony and natural gas dummy variables are 1.247762 (t-stat = 137.27) and 1.230578 (t-stat = 133.33), respectively. These values are consistent with our expectations. Balconies can increase the price of a condominium unit because it provides additional living space, as well as an ideal "observation post" for enjoying the view. Furthermore, natural gas is considered to be an important and preferred means of heating homes in Ottawa and so its presence should increase the price of a condominium unit.

Table 4 lists the estimated structure coefficients for Models (22) to (31). These additional coefficients estimates include the net geometric depreciation rate ($\hat{\delta}$), the bedroom dummy variables ($\hat{\tau}_k$), the bathroom dummy variables ($\hat{\varphi}_c$), the balcony dummy variable ($\hat{\pi}_2$) and the natural gas dummy variable ($\hat{\eta}_2$). A full list of the estimates from Model (31) are found in Appendix 4 (Subsection 8.4).

From Model (31), we also got an estimate for the average net geometric depreciation rate ($\hat{\delta}$), which was 0.023508 (t stat = 46.04), for the entire Q1 1996 to Q3 2009 period. It is slightly higher than the estimate used by Statistics Canada's Canadian System of Macroeconomic Accounts for deflating residential construction activity and for conducting productivity analysis.

However, it should be kept in mind that our estimated geometric depreciation rate of 2.4% per year may be subject to some downward bias for two reasons:

1. Capital expenditures on maintaining and renovating the structure are not taken into account in our model, so that we are estimating a net of capital expenditures depreciation rate rather than a gross depreciation rate,
2. Our model does not take into account the premature demolition of condominium buildings; that is, we only observe sales of surviving structures. This could be a source of bias. This problem can be addressed if information on the age of buildings

Table 4. Estimates (and t statistics) for structure variables added in Models 22 to 31.

Structure variable	Coefficient	Model 22	Model 24	Model 27	Model 31
Depreciation rate	$\hat{\delta}$	0.010636 (10.41)	0.03107 (52.52)	0.027315 (49.22)	0.023508 (46.04)
2 Bedrooms	$\hat{\tau}_2$	-	2.672505 (125.94)	2.025372 (71.94)	1.536801 (78.92)
3 or 4 Bedrooms	$\hat{\tau}_3$	-	3.171411 (65.36)	2.179728 (52.31)	1.506198 (57.6)
2 Bathrooms	$\hat{\varphi}_2$	-	-	1.328436 (77.88)	1.387985 (89.44)
3 Bathrooms	$\hat{\varphi}_3$	-	-	1.478348 (22.56)	1.540539 (30.65)
Balcony	$\hat{\pi}_2$	-	-	-	1.247762 (137.27)
Natural Gas	$\hat{\eta}_2$	-	-	-	1.230578 (133.33)
R square		0.6975	0.7764	0.7848	0.8052
Log Likelihood		-119,869	-118,364	-118,172	-117,677

when they are demolished is available. Diewert and Shimizu (2017) has addressed this problem. For commercial structures in Tokyo, they found that the demolition depreciation added an additional 2% per year to their estimated net depreciation rate obtained using the Builder's Model.

When determining the important factors that contribute to land value or the total price of a property, it is important to note that, though the magnitude of the difference in log-likelihood values between models may vary, the choice of variables included in our model are not impacted by the order in which they are included in the model. To further assess the robustness of the choice of our property characteristics, we estimated our models over a range of time periods, specifically 1996–2003, 1996–2004, 1996–2005, 1996–2006, 1996–2007, 1996–2008, and 2003–2009. A sample of the results from the time periods of 1996 to 2003 and 1996 to 2008 are compared to Models 21 and 31 in Appendix 5 (Subsection 8.5). Our results were similar for the key determinants of land across all time periods (as shown in Table 9 in Appendix 5). When the full Builder's Model was assessed across all time periods, the estimates were similar and the same choices in characteristics remained the same. However, there was a notable difference in the value of $\hat{\rho}$, the land distribution weight estimate. In Model 21, there is a near 50–50 split between equal and proportional distribution of land to a condo unit. This is the case across all time periods tested. When we include the structure components to complete the Builder's Model, the land distribution weight greatly favours equal distribution of land. In Model 31, $\hat{\rho}$ has a value of 0.10284 (t stat $= 6.12$) meaning that equal distribution of land has a roughly 90% weight. However, in the early time periods, $\hat{\rho}$ is negative, meaning that proportional distribution of land has a negative weight. In the later time periods, for example 1996 to 2008 or 2003 to 2009, $\hat{\rho}$ closely resembles the value we received in Model 31 of 10%. The main conclusion to this analysis is that a robust Builder's Model can be estimated with only equal distribution of land to a condo unit.

6. The Resulting Price Indices

Now that we have estimates for land prices, we can use them to construct land, structure and total property price indices. The main goal of this article is to present a sound methodology to create separate land and structure price indices. These indices are quite unique in Canada, specifically there is no other land price index for the city of Ottawa to confront our results with. Therefore, we will be constructing a total property price index by aggregating the land and structure components developed in this article. This allows us to confront our results with more common methods of calculating price indices.

6.1. Constructing Land, Structure and Total Condominium Sales Price Indices

When we are modeling the value of land, (α_t estimates represent the percentage change in land price due to the change in time from period 1 to period t, and we can use these estimates to create land price indices (IL_t) as follows:

$$IL_t = \alpha_t \times 100; \quad \text{t} = 1, \ldots, 55. \tag{33}$$

The structure price index is a normalization of the official Statistics Canada apartment construction cost index (PS_t). Since we indexed the price per square foot of structure with the ABCPI to get an approximate value of structure price for each period t, our structure price index (IS_t) is implicitly estimated by the ABCPI based to Q1 1996 = 100:

$$IS_t = \frac{PS_t}{PS_1} \times 100 = ABCPI_t; \quad \text{t} = 1, \ldots, 55. \tag{34}$$

We start by calculating our total property price index using a fixed base Laspeyres price index formula, which is a basket index and widely used by statistical agencies, and it will also be used in compiling the NCAPI of Statistics Canada. To use our method for decomposing property value into land and structure components for the national balance sheets of a country, one would need weighting information on the stock of condominium apartments in the country. This information is difficult to obtain and so the weights we use in this study are the value shares of land and structures.

First, we calculate the predicated values of land (LV_{tn}) and structures (SV_{tn}) for t = 1,. . .,55 and n = 1,. . .,N_t:

$$\begin{aligned} LV_{tn} = {} & \alpha t (\textstyle\sum_{i=1}^{22} \theta_i FSA_{tn,i})(1 + \gamma(H_{tn} - 1))(1 + \omega(TU_{tn} - 9))(\textstyle\sum_{j=1}^{4} \vartheta_j TH_{tn,j}) \\ & (\textstyle\sum_{m=1}^{4} \sigma_m EL_{tn,m}) \left[\rho \left(\frac{S_{tn}}{TS_{tn}} \right) + (1 - \rho) \left(\frac{1}{TU_{tn}} \right) \right] TL_{tn}; \end{aligned} \quad . \tag{35}$$

$$\begin{aligned} SV_{tn} = {} & (1.33) PS_t (1 - \delta)^{A_{tn}} (\textstyle\sum_{k=1}^{3} \tau_k BD_{tn,k})(\textstyle\sum_{c=1}^{3} \varphi_c BT_{tn,c}) \\ & (\textstyle\sum_{y=1}^{2} \pi_y BC_{tn,y})(\textstyle\sum_{z=1}^{2} \eta_z NG_{tn,z}) S_{tn}. \end{aligned} \quad . \tag{36}$$

In order to get total land and structure values for sales in period t, we sum the predicted values from Models (35) and (36) to get:

$$LV_t = \textstyle\sum_{n=1}^{N_t} LV_{tn}; \quad \text{t} = 1, \ldots, 55. \tag{37}$$

$$SV_t = \textstyle\sum_{n=1}^{N_t} SV_{tn}; \quad \text{t} = 1, \ldots, 55. \tag{38}$$

We define the total property value of condominium sales for period t, V_t, as the sum of the predicted values LV_t and SV_t:

$$V_t = LV_t + SV_t; \quad \text{t} = 1, \ldots, 55. \tag{39}$$

The fixed base Laspeyres index formula for period t can be written as follows:

$$I_t = IL_t \left(\frac{LV_1}{V_1} \right) + IS_t \left(\frac{SV_1}{V_1} \right); \quad \text{t} = 1, \ldots, 55. \tag{40}$$

The land, structure and fixed base Laspeyres (or total property) price indices for sales of condominium units are illustrated in Figure 1.

We can see that land prices have increased 4.42 fold between Q1 1996 and Q3 2009. From discussions with potential users of our land index, including the Consumer Price

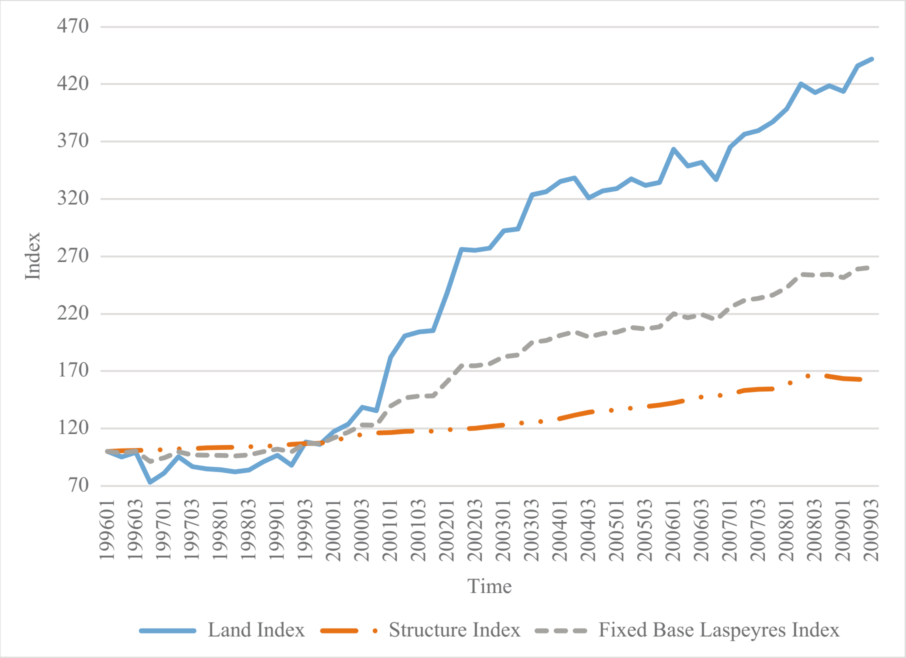


Fig. 1. Land, structure and fixed base Laspeyres price indices.

Index, the Canadian System of Macroeconomic Accounts, and other residential property price indices produced at Statistics Canada, these land results are deemed to be reasonable. This is the only condominium land price index of its kind in Canada, therefore, we cannot compare our results to any other land price index to legitimize them. However, other condominium price indices do exist that model the total property price of a unit. In Subsection 6.3, we will compare our total property price index to other indices that use different methods of calculation, but use the same variables as used in Model (31).

6.2. Land and Structure Value Shares and Alternative Total Property Price Indices

Before we go any further in comparing total property price indices, we have to decide which formula we will use to calculate our hedonically imputed index. The fixed base Laspeyres index shown in Figure 1 is misleading because over the 1996 to 2009 period, the land and structure value shares of condo sales change dramatically, as shown in Figure 2.

We can see that at our base period, Q1 1996, the structure component has a 65% share of the total value. However, as of Q1 2001, land takes over the majority share. This means that if we were to calculate a fixed base Paasche or a Fisher Index, the total property price index will look quite different. Figure 3 illustrates the difference between the fixed base Laspeyres, Paasche and Fisher indices calculated from our land and structure price and value estimates from Model (31).

Note that the fixed base Paasche and Fisher indices are higher than the Laspeyres. This is counter intuitive to most cases, where we see that the Paasche and Fisher indices are lower than the Laspeyres because of change in consumption patterns and weighting due to

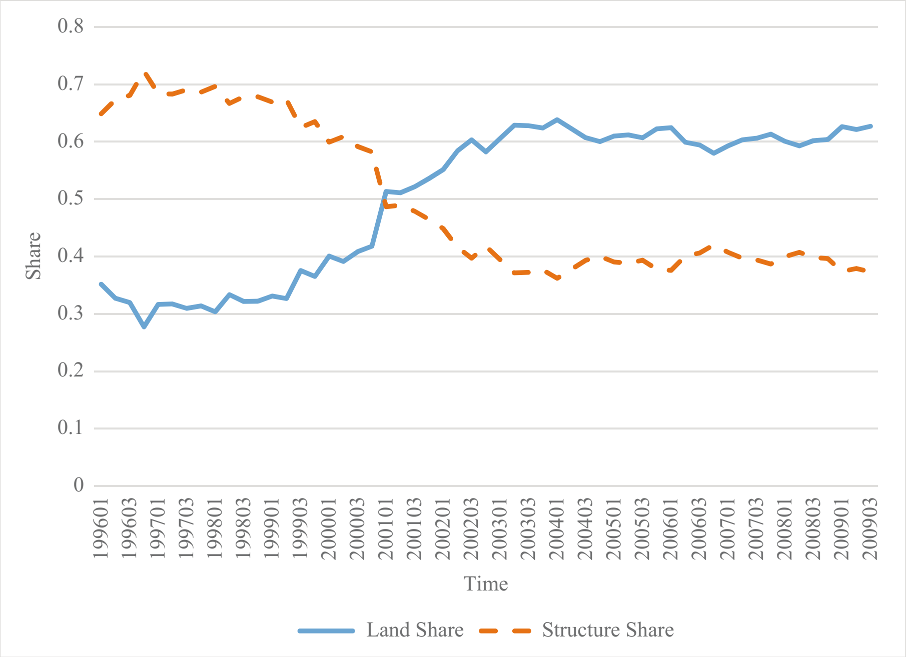


Fig. 2. Land and structure value shares over time.

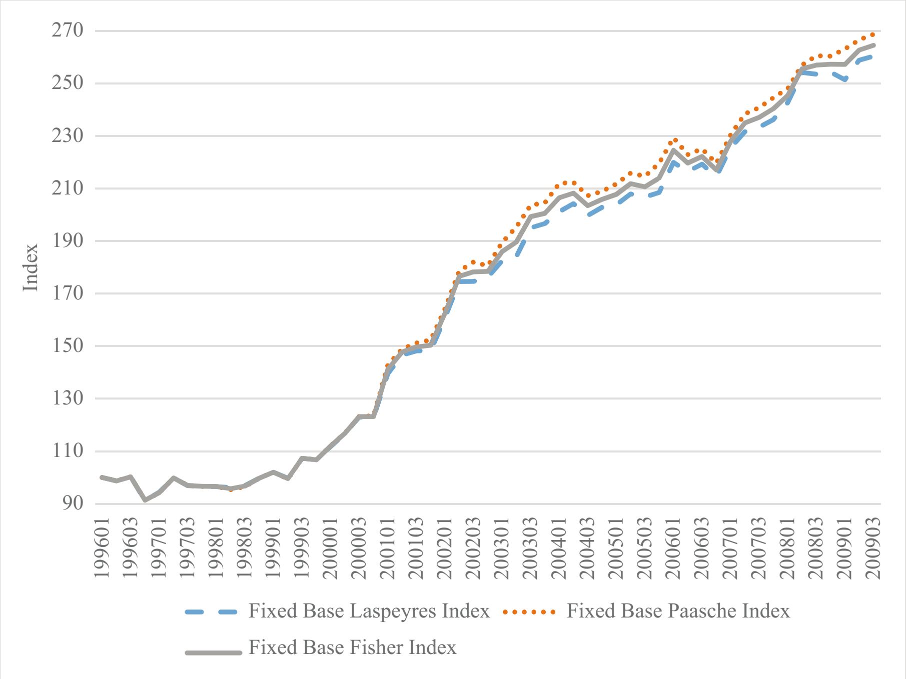


Fig. 3. Fixed base Laspeyres, Paasche and Fisher total property price indices.

preferences towards cheaper goods. However, starting in 2001, the land value share is dominant, meaning that the land value, which exhibits much more growth than the structure value, gets a higher weight.

Due to this phenomenon in weighting patterns, a chained index would display different results than its fixed counterpart. Figure 4 illustrates the differences between the chained and fixed base Laspeyres.

With weights in the chained Laspeyres being more timely than its fixed base counterpart, they are more representative for each comparison period, reflecting the changes in the land share over time.

In Figure 5. Chained Laspeyres versus Paasche price index, we see, using the chained methodology, that the Laspeyres index is higher than the Paasche, which follows traditional index theory (Diewert 2009).

Also, as predicted, the spread between the Laspeyres and Paasche is dramatically reduced, which is clearly shown in Figure 5. The chained indices more closely approximate each other than the fixed base indices, such that we chose not to illustrate the chained Fisher index in Figure 5 because it would be indistinguishable between its Laspeyres and Paasche counterparts. However, we also need to point out, with some bounces in the land prices, the chained indices could suffer a certain degree of chain drift.

6.3. Comparison With Other Total Property Indices

As mentioned in Subsection 6.1, we do not have any other official land price indices that can be compared to our land price index. However, we can compare our fixed base and

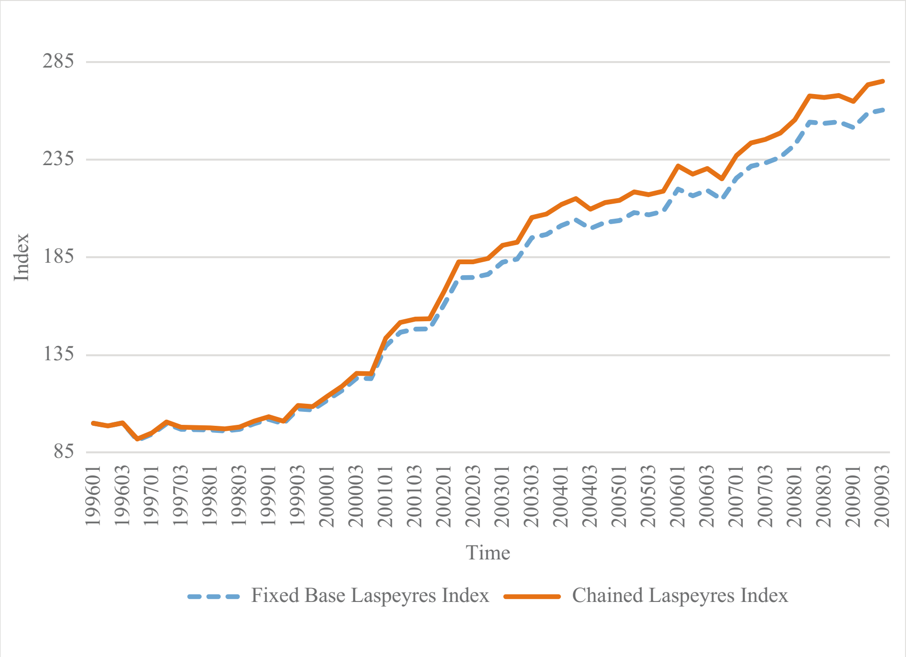


Fig. 4. Chained versus fixed base Laspeyres price indices.

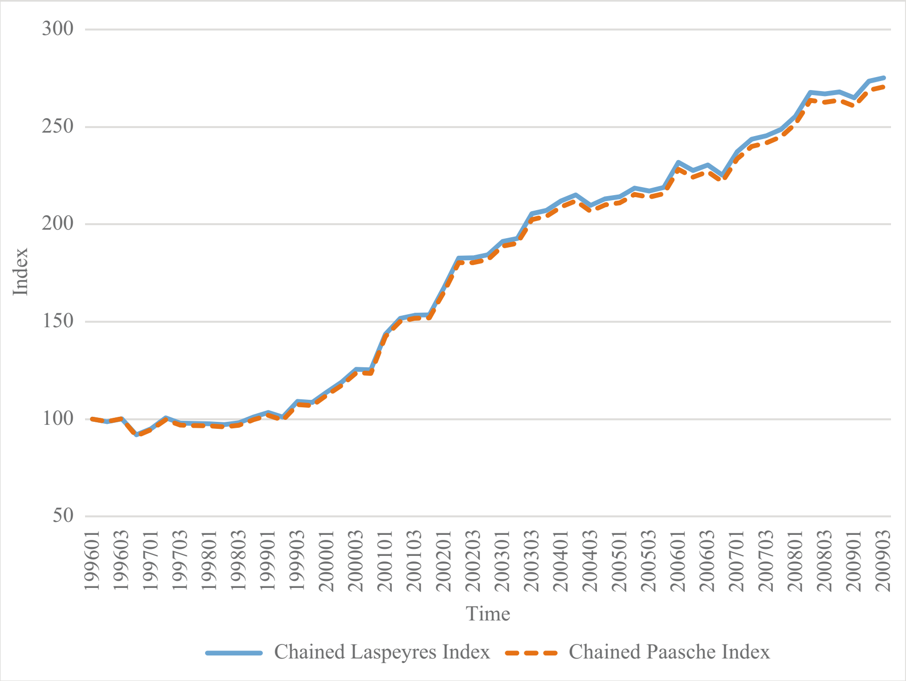


Fig. 5. Chained Laspeyres versus Paasche price index.

chained Fisher indices from Subsection 6.2 to total property indices calculated by other traditional methods, which are explained in detail in De Haan and Diewert (2013, 50–64), using the same explanatory variables.

First, we will compare our Fisher indices to three hedonic indices calculated by the following methods: the Pooled Time Dummy hedonic method, the Rolling Window Time Dummy hedonic method and the Hedonic Imputation approach. Hedonic methods have become a preferred method of constructing constant quality housing price indices, even though the data requirements are extensive and often expensive to obtain. All these hedonic methods regress the logarithm of selling price on selected characteristics for a certain time span, either a quarter, a year or all the periods under examination.

The characteristics that we included in these hedonic models are the same as were used in the Builder's Model (31). These characteristics reflect both quantitative and qualitative housing features that determine condominium prices. Figure 6 compares the different condo price indices for Ottawa using a traditional pooled time dummy hedonic regression.

The Pooled Time Dummy method runs a single regression on both characteristics variables and time dummy variables. It is very simple to apply in practice. The price index can be obtained directly from the estimated regression equation. The dependent variable is the logarithm of the unit's selling price and the overall price index is obtained by taking the exponential of the time dummy coefficients.

A practical problem associated with the hedonic regression model is the reassessment of the parameters with more recent data available. Figure 7 illustrates the Chained Fisher price index with a Rolling Window and a Hedonic Imputation Index.

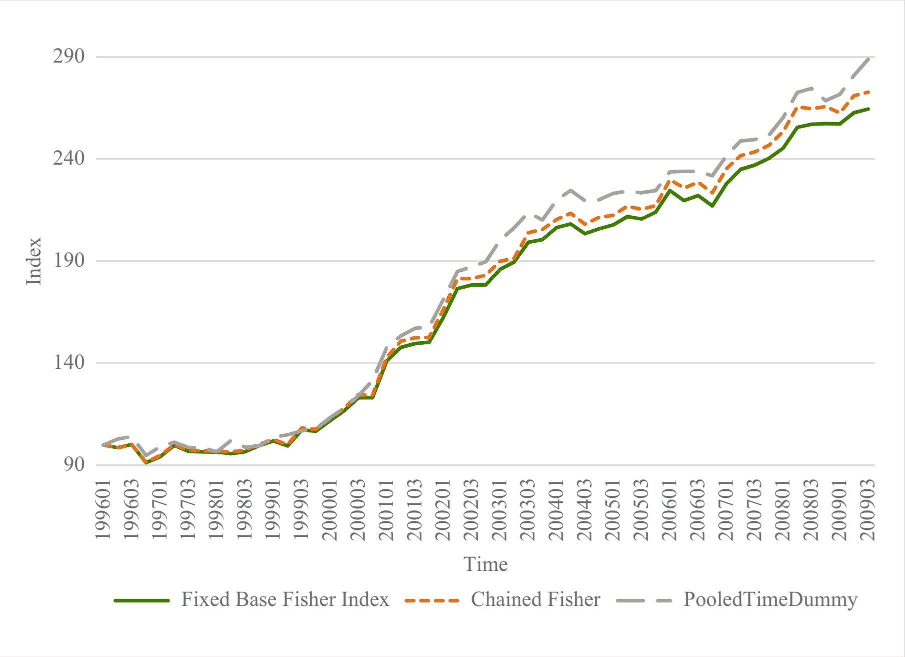


Fig. 6. Total property Fisher indices versus alternative hedonic regression base indices.

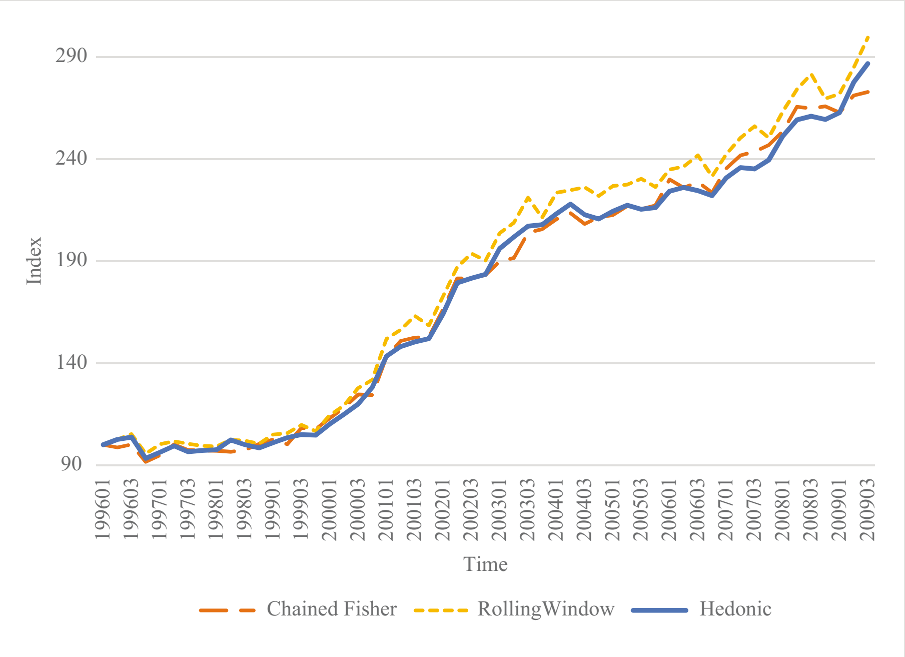


Fig. 7. Fixed base Fisher versus rolling window hedonic index and hedonic imputation Fisher index.

The Rolling Window approach is a simple solution to this problem. The Rolling Window Time Dummy method is similar to the Pooled Time Dummy method, with the difference that the Rolling Window Method runs a sequence of hedonic regressions for a fixed-window length, such as a year. This length of the window is determined when the model yields relatively robust estimates. We applied Rolling Window procedure with a length of five quarters. The advantage of this method over the Pooled Time Dummy method is that the Rolling Window method allows for gradual changes in consumer tastes or preferences over time.

In order to implement the Hedonic Imputation approach, a separate hedonic regression is run using the data for each period. In general, a set of fixed quantity of characteristics of a standard or matched model are chosen to impute the missing prices using the estimated coefficients from the hedonic regression model. Based on which time period the fixed characteristics belong to, the Laspeyres, Paasche and Fisher imputation indices can be estimated. The chained Fisher index calculated by using the Hedonic Imputation method (labeled Hedonic Imputation Fisher*)* is shown in Figure 7. We found that all three alternative hedonic indices generally approximate each other fairly closely.

From Figure 6 and Figure 7 we can see that the Fisher indices, calculated from our nonlinear Model (31), follow the same long-term trend as the Pooled Time Dummy, Rolling Window and Hedonic Imputation models. The fixed base Fisher and the chained Fisher indices exhibit a 1.69 and 1.76 fold increase, respectively, between Q1 1996 and Q3 2009. This is less than the total growth of the three alternative hedonic models. Specifically, the fixed base Fisher index has a 169% increase, the chained Fisher index has a 176% increase, the Pooled Time Dummy Index has a 188.9% increase, the Rolling Window Index has a 199.5% increase, and the Hedonic Imputation Method has a 186.7% increase. All four indices do have similar quarterly movements with an average growth rate of roughly 2% over the 14-year period.

The concern with using log-linear regression models such as the Pooled Time Dummy, Rolling Window and Hedonic Imputation models is that there can be multicollinearity between the variables, causing misleading coefficient estimates, which are then used to calculate the indices themselves. Therefore, we want to compare our Fisher Index to three indices using the following stratification methods: Mean Index and Median Index stratified by postal code and weighted by the sales in each quarter and the Median Index using stratification method proposed by Prasad and Richards (2006). These methods revolve around compiling a condominium price index using the mean or median price of each period. This methodology is simple and requires little information. However, this type of index has many disadvantages, such as it cannot fully account for quality change, and the compositional change of the housing stock will affect the price indices. Appropriate stratification can reduce bias caused by this compositional change.

Location is one of the natural stratification variables to use. We test the impact of using different fineness of classification schemes as the stratification indicator, with the finest neighbourhood variable in our data called district, the second finest called ward and the largest area in our data called FSA. Since condominium units are sold more frequently in certain areas and less frequently in the others, the alternative indices exhibit different price change patterns in different locations, which indicates that keeping the homogeneity of each cell is very important for the accuracy of the index. However, when the stratification scheme

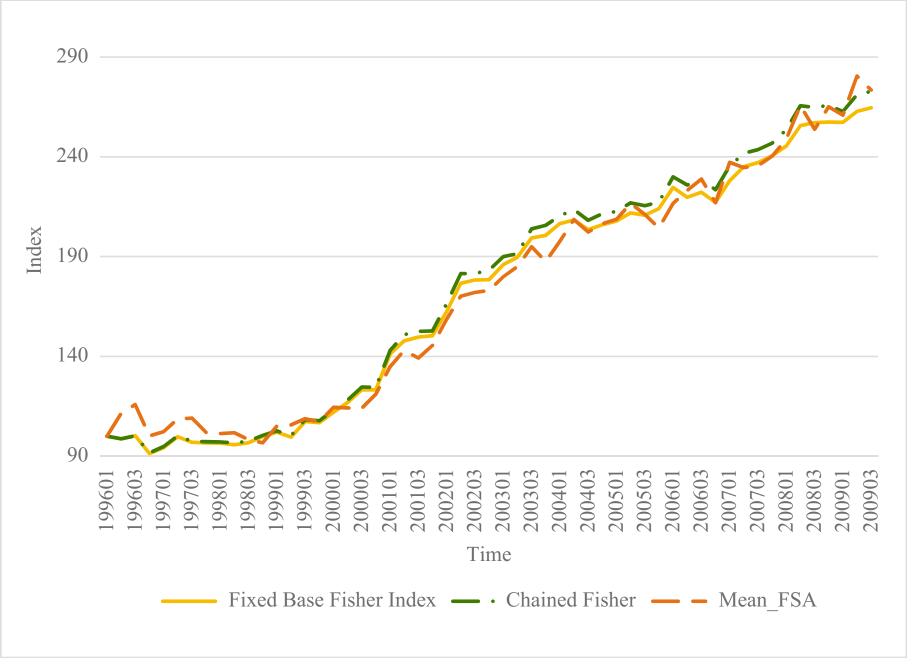


Fig. 8. Fisher indices versus mean index.

is too fine, empty cells will occur for some periods. If the classification scheme is very coarse, we cannot sufficiently control the homogeneity of the cell. The stratified price series reported in the article use FSA as the stratification variable. Figure 8, Figure 9, and Figure 10 illustrate the comparison between our Fisher indices and the Mean Index (Mean_FSA), Median Index (Med_FSA), and Median Index proposed by Prasad and Richards (Med (P&R)).

The period-to-period movements vary between the five indices. However, the long-term trends are similar across all five indices. It can be seen that there are more fluctuations in the Median index stratified by FSA, especially after the first quarter of 2006, than those in the other four price series. This might be a result of using sales as the weight to aggregate indices across different FSA.

7. Conclusion

The most important conclusion from this study is that we now have a method to create land price indices for condominium units. Moreover, our estimated structure price index can be harmonized with current structure price indices that are used in the System of National Accounts. Condominium land and building characteristics data are difficult to find in Canada and attaining these data is a hurdle in putting the Builder's Model into practice. If the required information is obtained, we could apply this method to fill the missing gaps in the production of Statistics Canada's New Condominium Apartment Price Index and future residential property price indices. Though we cannot fully determine the accuracy of our land index by comparing to other sources, because no such sources exist, the similarities between the Fisher indices created from our Builder's Model and other hedonic and stratification methods, are promising for our proposed method of index calculation.

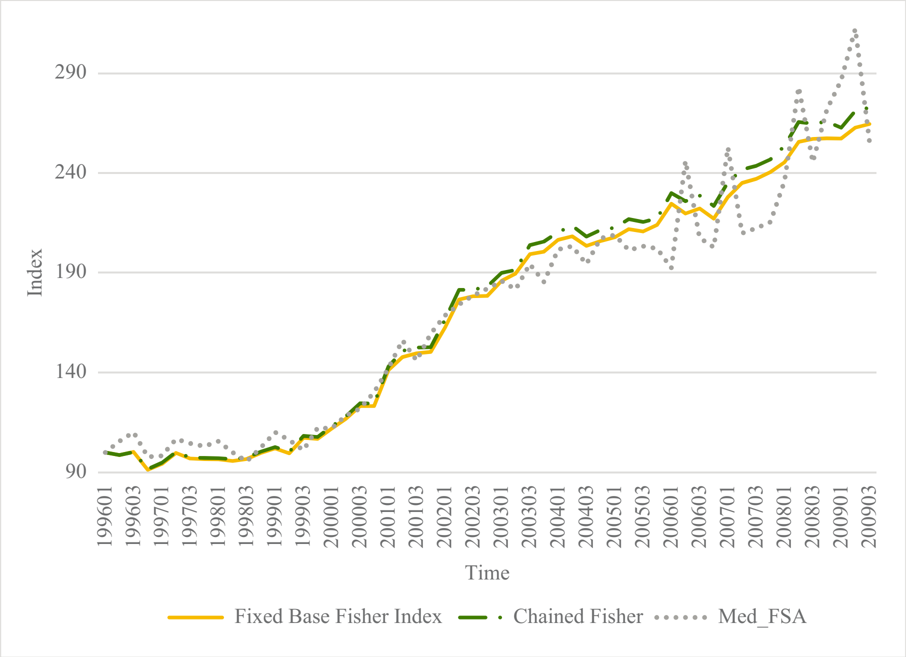


Fig. 9. Fixed base and chained Fisher indices versus median FSA index.

Through our modeling, we narrowed down the significant determinants of land prices to include location (determined by FSA), unit height, number of units in the building, building height and excess land. Our measurement of location by using FSA dummy

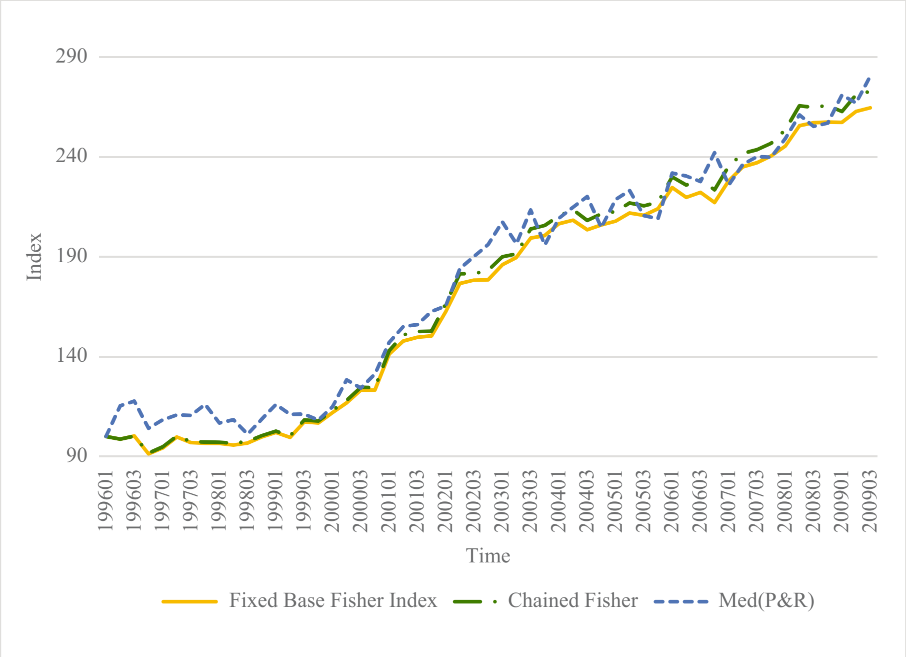


Fig. 10. Fixed base and chained Fisher indices versus median (P&R) index.

variables is rather discrete. To improve our assessment of location, including neighborhood characteristics, this model will need to be applied to other geographic areas.

We also identified structure quality adjustment variables, such as the number of bedrooms, number of bathrooms, the presence of a balcony and natural gas heating in a condominium unit that impact price. Many other variables, such as dens, hardwood floors, condo fees and on-suite bathrooms were tested in our Builder's Model that appeared to have little impact. Comparing with the variables included in the model of Diewert and Shimizu (2016), we believe that the city characteristics will also have an impact on determining the choice of variables added to the Builder's Model. For instance, due to the long winters in Ottawa, the means of heating is an important feature for determining the price of condominium units. All these findings could be helpful for designing a survey to effectively collect required information at a minimum cost.

Lastly, we determined a net geometric depreciation rate of 2.4% for the Q1 1996 to Q3 2009 period. This value is slightly larger than that currently used by the Canadian System of National Accounts. As noted earlier, demolition depreciation is neglected in our model and so a geometric depreciation rate of 2.4% should be regarded as a lower bound on the overall depreciation rate. This exercise highlights that not only can the Builder's Model provide a land price index for the National Accounts, but it can also provide additional beneficial statistics for other parts of the System of National Accounts.

8. Appendix

8.1. Appendix 1: Regression Results Using 20, 25 and 30% Communal Space in Model 22

Table 5. Regression results using 20, 25 and 30% communal space in model 22.

	20%		25%		30%	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\theta}_1$	92.3106	6.14	87.9885	6.07	82.9916	5.98
$\hat{\theta}_2$	109.845	8.69	105.601	8.54	100.551	8.35
$\hat{\theta}_3$	115.028	8.48	110.461	8.34	105.055	8.16
$\hat{\theta}_4$	70.1347	8.45	67.0257	8.31	63.4037	8.13
$\hat{\theta}_5$	99.8556	8.67	96.0765	8.51	91.619	8.33
$\hat{\theta}_6$	86.4793	8.59	83.3044	8.44	79.5637	8.26
$\hat{\theta}_7$	126.78	8.62	122.255	8.47	116.863	8.29
$\hat{\theta}_8$	193.445	8.76	186.962	8.6	179.208	8.41
$\hat{\theta}_9$	163.113	8.73	157.525	8.57	150.892	8.39
$\hat{\theta}_{10}$	224.005	8.66	216.273	8.51	207.065	8.33
$\hat{\theta}_{11}$	92.7051	8.24	88.5551	8.1	83.6586	7.93
$\hat{\theta}_{12}$	110.906	8.67	106.395	8.52	101.075	8.33
$\hat{\theta}_{13}$	150.972	8.72	145.54	8.57	139.11	8.38
$\hat{\theta}_{14}$	280.319	8.69	271.132	8.54	260.187	8.35
$\hat{\theta}_{15}$	160.981	8.63	155.352	8.48	148.691	8.3
$\hat{\theta}_{16}$	101.275	8.67	97.4698	8.52	92.9864	8.33
$\hat{\theta}_{17}$	100.839	8.5	96.7042	8.35	91.8581	8.17
$\hat{\theta}_{18}$	119.712	8.69	115.267	8.53	110.042	8.35
$\hat{\theta}_{19}$	110.827	8.65	106.175	8.5	100.758	8.31

Table 5. Continued

	20%		25%		30%	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\theta}_{20}$	133.896	7.66	128.443	7.55	122.009	7.41
$\hat{\theta}_{21}$	206.157	8.76	199.379	8.61	191.291	8.42
$\hat{\theta}_{22}$	148.594	7.28	142.856	7.19	136.031	7.07
$\hat{\alpha}_{2}$	1.04534	7.14	1.04778	7.01	1.05093	6.85
$\hat{\alpha}_{3}$	1.13726	7.1	1.14015	6.98	1.14385	6.83
$\hat{\alpha}_{4}$	0.87045	5.41	0.86845	5.3	0.86598	5.17
$\hat{\alpha}_{5}$	0.83845	8	0.83843	7.83	0.83853	7.63
$\hat{\alpha}_{6}$	1.03381	7.34	1.03479	7.2	1.03622	7.03
$\hat{\alpha}_{7}$	0.96328	7.1	0.96336	6.96	0.96362	6.78
$\hat{\alpha}_{8}$	0.92834	6.32	0.928	6.19	0.92777	6.02
$\hat{\alpha}_{9}$	0.93718	6.77	0.93608	6.63	0.93483	6.46
$\hat{\alpha}_{10}$	0.79378	8.5	0.79269	8.34	0.79131	8.15
$\hat{\alpha}_{11}$	0.8809	6.68	0.87994	6.53	0.87897	6.35
$\hat{\alpha}_{12}$	0.96838	5.28	0.96732	5.18	0.96622	5.07
$\hat{\alpha}_{13}$	1.00982	7.23	1.01083	7.09	1.0124	6.92
$\hat{\alpha}_{14}$	0.88547	8.55	0.88524	8.4	0.88503	8.2
$\hat{\alpha}_{15}$	1.02688	7.69	1.02897	7.54	1.032	7.36
$\hat{\alpha}_{16}$	1.03449	6.73	1.03658	6.62	1.03962	6.48
$\hat{\alpha}_{17}$	1.08156	7.49	1.08566	7.36	1.09134	7.2
$\hat{\alpha}_{18}$	1.13656	8.09	1.14082	7.94	1.14671	7.76
$\hat{\alpha}_{19}$	1.20013	7.43	1.20469	7.29	1.21098	7.13
$\hat{\alpha}_{20}$	1.13941	8.59	1.14462	8.44	1.15142	8.25
$\hat{\alpha}_{21}$	1.36364	8.66	1.37589	8.51	1.39192	8.32
$\hat{\alpha}_{22}$	1.54022	8.45	1.55556	8.3	1.57576	8.12
$\hat{\alpha}_{23}$	1.53205	8.29	1.54807	8.15	1.56913	7.97
$\hat{\alpha}_{24}$	1.51548	8.52	1.53228	8.37	1.55434	8.19
$\hat{\alpha}_{25}$	1.76457	8.61	1.7875	8.46	1.81741	8.27
$\hat{\alpha}_{26}$	1.9749	8.61	2.00419	8.46	2.04237	8.28
$\hat{\alpha}_{27}$	1.9082	8.57	1.93553	8.42	1.97113	8.24
$\hat{\alpha}_{28}$	1.96541	8.49	1.99381	8.35	2.0308	8.17
$\hat{\alpha}_{29}$	2.04157	8.58	2.07241	8.44	2.11263	8.26
$\hat{\alpha}_{30}$	1.94756	8.75	1.97596	8.6	2.01269	8.41
$\hat{\alpha}_{31}$	2.1943	8.66	2.22967	8.51	2.27582	8.33
$\hat{\alpha}_{32}$	2.19107	8.58	2.22528	8.43	2.26984	8.25
$\hat{\alpha}_{33}$	2.24306	8.7	2.27907	8.55	2.32605	8.36
$\hat{\alpha}_{34}$	2.29486	8.73	2.33091	8.58	2.37799	8.39
$\hat{\alpha}_{35}$	2.21382	8.69	2.24682	8.54	2.28992	8.35
$\hat{\alpha}_{36}$	2.25073	8.64	2.28557	8.49	2.33107	8.3
$\hat{\alpha}_{37}$	2.26801	8.72	2.30224	8.56	2.34699	8.38
$\hat{\alpha}_{38}$	2.29853	8.76	2.3341	8.6	2.38054	8.41
$\hat{\alpha}_{39}$	2.31549	8.71	2.35055	8.55	2.39645	8.37
$\hat{\alpha}_{40}$	2.32024	8.57	2.35537	8.42	2.40132	8.24
$\hat{\alpha}_{41}$	2.47301	8.74	2.5125	8.58	2.56415	8.39
$\hat{\alpha}_{42}$	2.46739	8.76	2.50482	8.6	2.55387	8.41
$\hat{\alpha}_{43}$	2.43186	8.75	2.46879	8.6	2.51717	8.41
$\hat{\alpha}_{44}$	2.39763	8.71	2.43322	8.55	2.47982	8.37

Table 5. Continued

	20%		25%		30%	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\alpha}_{45}$	2.5445	8.76	2.5827	8.61	2.63271	8.42
$\hat{\alpha}_{46}$	2.63324	8.78	2.67442	8.62	2.72824	8.43
$\hat{\alpha}_{47}$	2.65067	8.76	2.69218	8.61	2.74652	8.42
$\hat{\alpha}_{48}$	2.67028	8.74	2.71348	8.59	2.77002	8.4
$\hat{\alpha}_{49}$	2.79316	8.76	2.83799	8.6	2.89673	8.42
α_{50}	2.96019	8.79	3.00787	8.63	3.07045	8.44
$\hat{\alpha}_{51}$	2.83087	8.77	2.87605	8.61	2.93536	8.42
$\hat{\alpha}_{52}$	2.95191	8.77	2.99937	8.61	3.06159	8.42
$\hat{\alpha}_{53}$	2.83705	8.78	2.88212	8.62	2.94106	8.43
$\hat{\alpha}_{54}$	3.01506	8.79	3.06634	8.63	3.13345	8.44
$\hat{\alpha}_{55}$	2.93895	8.79	2.9895	8.63	3.05567	8.44
$\hat{\gamma}$	0.00834	13.51	0.00847	13.64	0.00864	13.8
$\hat{\omega}$	0.01004	39.23	0.01001	39	0.00995	38.71
$\hat{\vartheta}_2$	1.15636	106.96	1.15861	106.51	1.16203	105.98
$\hat{\vartheta}_3$	1.4345	77.2	1.4374	76.88	1.4423	76.52
$\hat{\vartheta}_4$	1.68029	69.79	1.68332	69.47	1.68884	69.1
$\hat{\sigma}_2$	0.54132	124.2	0.54203	123.74	0.54304	123.21
$\hat{\sigma}_3$	0.27106	81.76	0.2718	81.31	0.27283	80.78
$\hat{\sigma}_4$	0.18369	55.92	0.18394	55.61	0.18437	55.26
$\hat{\rho}$	0.50292	57.1	0.49601	55.63	0.48743	53.84
$\hat{\delta}$	0.01035	9.5	0.01064	10.41	0.01105	11.63

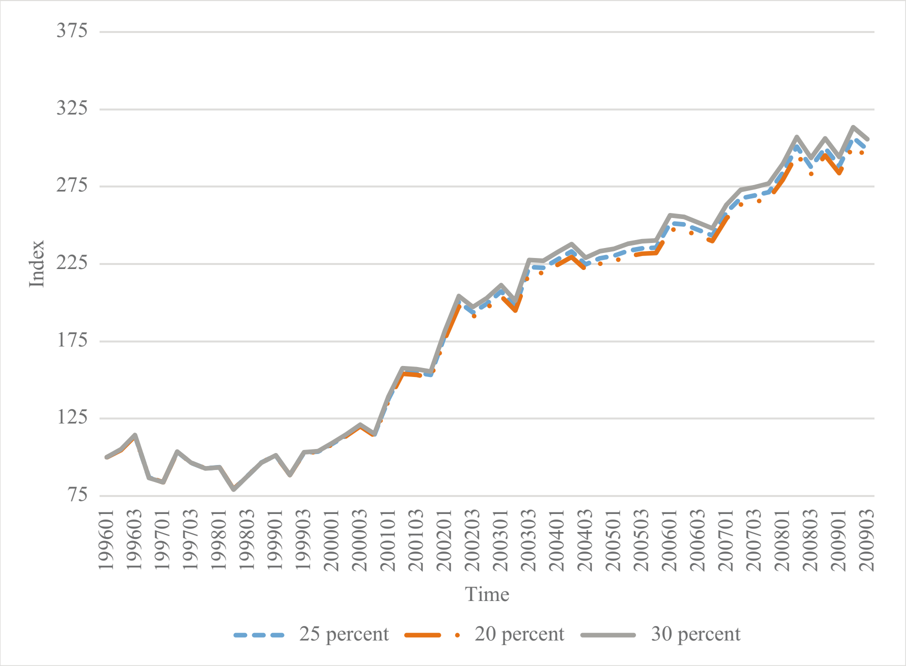


Fig. 11. Index results using 20, 25 and 30% communal space in Model 21.

8.2. Appendix 2: Estimates of Model 5

Table 6. Estimates of model 5.

Coefficient	Estimate	T stat	Coefficient	Estimate	T stat	Coefficient	Estimate	T stat
$\hat{\alpha}_2$	-71.219	-11.28	$\hat{\alpha}_{21}$	-34.208	-11.77	$\hat{\alpha}_{39}$	9.16141	1.58
$\hat{\alpha}_3$	-89.113	-7.19	$\hat{\alpha}_{22}$	-46.209	-6.09	$\hat{\alpha}_{40}$	13.3754	1.8
$\hat{\alpha}_4$	-91.878	-10.14	$\hat{\alpha}_{23}$	-39.033	-4.69	$\hat{\alpha}_{41}$	18.5442	3.37
$\hat{\alpha}_5$	-68.129	-29.66	$\hat{\alpha}_{24}$	-35.446	-3.83	$\hat{\alpha}_{42}$	9.60093	2.18
$\hat{\alpha}_6$	-81.54	-11.7	$\hat{\alpha}_{25}$	-18.88	-2.7	$\hat{\alpha}_{43}$	-4.1572	-0.84
$\hat{\alpha}_7$	-93.266	-10.65	$\hat{\alpha}_{26}$	-1.7848	-0.31	$\hat{\alpha}_{44}$	-12.04	-1.86
$\hat{\alpha}_8$	-86.648	-9.74	$\hat{\alpha}_{27}$	-3.8512	-0.57	$\hat{\alpha}_{45}$	9.30573	1.68
$\hat{\alpha}_9$	-87.328	-10.03	$\hat{\alpha}_{28}$	-6.658	-0.68	$\hat{\alpha}_{46}$	13.6358	3.06
$\hat{\alpha}_{10}$	-65.411	-30.05	$\hat{\alpha}_{29}$	10.4451	1.18	$\hat{\alpha}_{47}$	9.42861	2.13
$\hat{\alpha}_{11}$	-86.905	-9.85	$\hat{\alpha}_{30}$	12.0305	2.28	$\hat{\alpha}_{48}$	11.3436	2.23
$\hat{\alpha}_{12}$	-84.116	-7.99	$\hat{\alpha}_{31}$	15.3973	2.54	$\hat{\alpha}_{49}$	18.7812	4.13
$\hat{\alpha}_{13}$	-81.533	-11.8	$\hat{\alpha}_{32}$	9.33577	1.16	α_{50}	20.9311	5.4
$\hat{\alpha}_{14}$	-66.035	-33.19	$\hat{\alpha}_{33}$	17.46	2.56	$\hat{\alpha}_{51}$	5.79297	1.18
$\hat{\alpha}_{15}$	-84.299	-11.72	$\hat{\alpha}_{34}$	19.0803	3.24	$\hat{\alpha}_{52}$	15.821	3.11
$\hat{\alpha}_{16}$	-76.347	-8.72	$\hat{\alpha}_{35}$	3.68198	0.63	$\hat{\alpha}_{53}$	25.0618	5.47
$\hat{\alpha}_{17}$	-75.084	-9.68	$\hat{\alpha}_{36}$	10.1405	1.54	$\hat{\alpha}_{54}$	41.1904	11.19
$\hat{\alpha}_{18}$	-72.835	-11.84	$\hat{\alpha}_{37}$	11.235	1.66	$\hat{\alpha}_{55}$	38.7612	9.21
$\hat{\alpha}_{19}$	-78.822	-9.88	$\hat{\alpha}_{38}$	15.3105	2.92	β	5.02528	330.25
$\hat{\alpha}_{20}$	-58.015	-21.86						

8.3. Appendix 3: Estimates of Model 21

Table 7. Estimates of model 21.

Coefficient	Estimate	T stat	Coefficient	Estimate	T stat	Coefficient	Estimate	T stat
$\hat{\theta}_1$	100.267	6.53	$\hat{\alpha}_9$	0.95655	7.12	$\hat{\alpha}_{37}$	2.28565	9.09
$\hat{\theta}_2$	150.452	7.58	$\hat{\alpha}_{10}$	0.81215	8.9	$\hat{\alpha}_{38}$	2.31307	9.14
$\hat{\theta}_3$	112.7	9.09	$\hat{\alpha}_{11}$	0.90267	7.01	$\hat{\alpha}_{39}$	2.32874	9.08
$\hat{\theta}_4$	118.564	8.85	$\hat{\alpha}_{12}$	0.99056	5.65	$\hat{\alpha}_{40}$	2.33858	8.94
$\hat{\theta}_5$	75.7725	8.87	$\hat{\alpha}_{13}$	1.03096	7.56	$\hat{\alpha}_{41}$	2.48442	9.11
$\hat{\theta}_6$	103.868	9.07	$\hat{\alpha}_{14}$	0.90194	8.95	$\hat{\alpha}_{42}$	2.4825	9.14
$\hat{\theta}_7$	90.4379	8.99	$\hat{\alpha}_{15}$	1.04931	8.06	$\hat{\alpha}_{43}$	2.44861	9.13
$\hat{\theta}_8$	131.426	9.02	$\hat{\alpha}_{16}$	1.05438	7.11	$\hat{\alpha}_{44}$	2.41217	9.08
$\hat{\theta}_9$	197.067	9.15	$\hat{\alpha}_{17}$	1.10472	7.86	$\hat{\alpha}_{45}$	2.56122	9.14
$\hat{\theta}_{10}$	169.31	9.12	$\hat{\alpha}_{18}$	1.16242	8.45	$\hat{\alpha}_{46}$	2.64082	9.16
$\hat{\theta}_{11}$	229.093	9.05	$\hat{\alpha}_{19}$	1.22093	7.77	$\hat{\alpha}_{47}$	2.66495	9.14
$\hat{\theta}_{12}$	95.6961	8.63	$\hat{\alpha}_{20}$	1.15979	8.99	$\hat{\alpha}_{48}$	2.68254	9.12
$\hat{\theta}_{13}$	115.48	9.06	$\hat{\alpha}_{21}$	1.37772	9.05	$\hat{\alpha}_{49}$	2.81145	9.14
$\hat{\theta}_{14}$	155.427	9.11	$\hat{\alpha}_{22}$	1.55776	8.83	$\hat{\alpha}_{50}$	2.98276	9.17
$\hat{\theta}_{15}$	281.362	9.06	$\hat{\alpha}_{23}$	1.5496	8.65	$\hat{\alpha}_{51}$	2.84972	9.15
$\hat{\theta}_{16}$	165.822	9.02	$\hat{\alpha}_{24}$	1.53571	8.89	$\hat{\alpha}_{52}$	2.96522	9.14
$\hat{\theta}_{17}$	105.691	9.08	$\hat{\alpha}_{25}$	1.77596	8.98	$\hat{\alpha}_{53}$	2.85257	9.16
$\hat{\theta}_{18}$	105.958	8.92	$\hat{\alpha}_{26}$	1.98367	8.98	$\hat{\alpha}_{54}$	3.02278	9.17
$\hat{\theta}_{19}$	123.421	9.09	$\hat{\alpha}_{27}$	1.91648	8.94	$\hat{\alpha}_{55}$	2.96293	9.17
$\hat{\theta}_{20}$	113.735	9.05	$\hat{\alpha}_{28}$	1.97515	8.86	$\hat{\gamma}$	0.00821	13.71
$\hat{\theta}_{21}$	138.525	8.03	$\hat{\alpha}_{29}$	2.05311	8.96	$\hat{\omega}$	0.00958	40.04
$\hat{\theta}_{22}$	211.818	9.16	$\hat{\alpha}_{30}$	1.96029	9.13	$\hat{\vartheta}_2$	1.1668	109.17
$\hat{\alpha}_2$	1.04855	7.38	$\hat{\alpha}_{31}$	2.20211	9.04	$\hat{\vartheta}_3$	1.45254	80.2
$\hat{\alpha}_3$	1.14032	7.43	$\hat{\alpha}_{32}$	2.19385	8.94	$\hat{\vartheta}_4$	1.71222	72.32
$\hat{\alpha}_4$	0.88626	5.73	$\hat{\alpha}_{33}$	2.25753	9.07	$\hat{\sigma}_2$	0.55125	131.43
$\hat{\alpha}_5$	0.85509	8.44	$\hat{\alpha}_{34}$	2.30651	9.11	$\hat{\sigma}_3$	0.27964	86.2
$\hat{\alpha}_6$	1.04669	7.65	$\hat{\alpha}_{35}$	2.22761	9.06	$\hat{\sigma}_4$	0.18943	59.74
$\hat{\alpha}_7$	0.97601	7.38	$\hat{\alpha}_{36}$	2.26332	9.01	$\hat{\rho}$	0.51334	63.49
$\hat{\alpha}_8$	0.94781	6.65						

8.4. Appendix 4: Estimates of Model 31

Table 8. Estimates of model 31.

Coefficient	Estimate	T stat	Coefficient	Estimate	T stat	Coefficient	Estimate	T stat
$\hat{\theta}_1$	18.9504	5.14	$\hat{\alpha}_{11}$	0.83732	4.27	$\hat{\alpha}_{42}$	3.48696	6.28
$\hat{\theta}_2$	51.5232	6.01	$\hat{\alpha}_{12}$	0.90912	3.97	$\hat{\alpha}_{43}$	3.51955	6.28
$\hat{\theta}_3$	37.792	6.22	$\hat{\alpha}_{13}$	0.96641	4.94	$\hat{\alpha}_{44}$	3.3684	6.25
$\hat{\theta}_4$	39.7257	6.07	$\hat{\alpha}_{14}$	0.8793	5.94	$\hat{\alpha}_{45}$	3.65136	6.28
$\hat{\theta}_5$	22.9675	6.11	$\hat{\alpha}_{15}$	1.08107	5.38	$\hat{\alpha}_{46}$	3.76422	6.29
$\hat{\theta}_6$	35.2053	6.2	$\hat{\alpha}_{16}$	1.06128	4.85	$\hat{\alpha}_{47}$	3.7964	6.28
$\hat{\theta}_7$	36.4712	6.21	$\hat{\alpha}_{17}$	1.17468	5.72	$\hat{\alpha}_{48}$	3.87238	6.27
$\hat{\theta}_8$	51.3282	6.22	$\hat{\alpha}_{18}$	1.23711	5.87	$\hat{\alpha}_{49}$	3.98489	6.29
$\hat{\theta}_9$	74.1072	6.25	$\hat{\alpha}_{19}$	1.38538	5.45	$\hat{\alpha}_{50}$	4.20267	6.29
$\hat{\theta}_{10}$	65.5811	6.25	$\hat{\alpha}_{20}$	1.35506	6.15	$\hat{\alpha}_{51}$	4.12712	6.28
$\hat{\theta}_{11}$	77.1784	6.21	$\hat{\alpha}_{21}$	1.82033	6.2	$\hat{\alpha}_{52}$	4.18693	6.28
$\hat{\theta}_{12}$	29.2054	6.01	$\hat{\alpha}_{22}$	2.00595	6.04	$\hat{\alpha}_{53}$	4.1375	6.29
$\hat{\theta}_{13}$	38.9776	6.17	$\hat{\alpha}_{23}$	2.0415	6.02	$\hat{\alpha}_{54}$	4.3602	6.3
$\hat{\theta}_{14}$	60.2137	6.24	$\hat{\alpha}_{24}$	2.052	6.1	$\hat{\alpha}_{55}$	4.4187	6.3
$\hat{\theta}_{15}$	96.2797	6.2	$\hat{\alpha}_{25}$	2.38189	6.22	$\hat{\gamma}$	0.01035	14.87
$\hat{\theta}_{16}$	58.3052	6.2	$\hat{\alpha}_{26}$	2.76026	6.21	$\hat{\omega}$	0.01202	37.52
$\hat{\theta}_{17}$	37.3536	6.22	$\hat{\alpha}_{27}$	2.75262	6.2	$\hat{\vartheta}_2$	1.11502	95.13
$\hat{\theta}_{18}$	37.2665	6.14	$\hat{\alpha}_{28}$	2.77131	6.13	$\hat{\vartheta}_3$	1.42347	72.63
$\hat{\theta}_{19}$	45.4616	6.21	$\hat{\alpha}_{29}$	2.92292	6.2	$\hat{\vartheta}_4$	1.58045	64.35
$\hat{\theta}_{20}$	38.1906	6.17	$\hat{\alpha}_{30}$	2.93856	6.28	$\hat{\sigma}_2$	0.56898	120.51
$\hat{\theta}_{21}$	44.821	5.89	$\hat{\alpha}_{31}$	3.23666	6.24	$\hat{\sigma}_3$	0.30705	72.27
$\hat{\theta}_{22}$	81.1283	6.26	$\hat{\alpha}_{32}$	3.26326	6.21	$\hat{\sigma}_4$	0.20825	55.63
$\hat{\alpha}_2$	0.95139	4.59	$\hat{\alpha}_{33}$	3.35212	6.27	$\hat{\rho}$	0.10284	6.12
$\hat{\alpha}_3$	0.99103	4.78	$\hat{\alpha}_{34}$	3.38355	6.28	$\hat{\delta}$	0.02351	46.04
$\hat{\alpha}_4$	0.7312	3.73	$\hat{\alpha}_{35}$	3.2087	6.26	$\hat{\tau}_2$	1.5368	78.92
$\hat{\alpha}_5$	0.8116	5.01	$\hat{\alpha}_{36}$	3.27093	6.23	$\hat{\tau}_3$	1.5062	57.6
$\hat{\alpha}_6$	0.95248	5	$\hat{\alpha}_{37}$	3.29137	6.27	$\hat{\varphi}_2$	1.38799	89.44
$\hat{\alpha}_7$	0.86708	4.99	$\hat{\alpha}_{38}$	3.37566	6.28	$\hat{\varphi}_3$	1.54054	30.65
$\hat{\alpha}_8$	0.84822	4.43	$\hat{\alpha}_{39}$	3.31833	6.26	$\hat{\varphi}_2$	1.24776	137.27
$\hat{\alpha}_9$	0.83986	4.65	$\hat{\alpha}_{40}$	3.34384	6.21	$\hat{\eta}_2$	1.23058	133.3
$\hat{\alpha}_{10}$	0.82253	5.94	$\hat{\alpha}_{41}$	3.63297	6.27			

8.5. Appendix 5: Robustness Testing

Table 9. Determinants of land value over selected time periods.

	1996–2009		1996–2003		1996–2008	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\theta}_1$	100.267	6.53	79.9762	7.03	96.2423	7.24
$\hat{\theta}_2$	150.452	7.58	134.34	6.96	149.643	8.65
$\hat{\theta}_3$	112.7	9.09	97.7052	17.72	113.125	10.55
$\hat{\theta}_4$	118.564	8.85	110.558	16.23	120.625	10.28
$\hat{\theta}_5$	75.7725	8.87	56.7741	13.95	74.7073	10.09
$\hat{\theta}_6$	103.868	9.07	87.8267	18.37	102.814	10.56
$\hat{\theta}_7$	90.4379	8.99	77.1447	16.72	91.0148	10.46
$\hat{\theta}_8$	131.426	9.02	121.379	16.26	129.148	10.43
$\hat{\theta}_9$	197.067	9.15	168.239	19.1	200.93	10.67
$\hat{\theta}_{10}$	169.31	9.12	138.186	18.49	161.353	10.62
$\hat{\theta}_{11}$	229.093	9.05	223.235	15.58	228.791	10.5
$\hat{\theta}_{12}$	95.6961	8.63	89.0804	14.83	94.7354	9.97
$\hat{\theta}_{13}$	115.48	9.06	105.145	17.42	116.302	10.52
$\hat{\theta}_{14}$	155.427	9.11	120.623	18.15	154.305	10.62
$\hat{\theta}_{15}$	281.362	9.06	-	-	195.441	8.21
$\hat{\theta}_{16}$	165.822	9.02	143.66	16.69	165.004	10.44
$\hat{\theta}_{17}$	105.691	9.08	98.2246	17.98	107.777	10.55
$\hat{\theta}_{18}$	105.958	8.92	83.7896	16.98	103.171	10.33
$\hat{\theta}_{19}$	123.421	9.09	123.221	18.06	128.488	10.55
$\hat{\theta}_{20}$	113.735	9.05	106.728	16.39	158.293	10.43
$\hat{\theta}_{21}$	138.525	8.03	130.076	13.41	142.312	9.11
$\hat{\theta}_{22}$	211.818	9.16	175.516	19.39	211.934	10.69
$\hat{\alpha}_2$	1.04855	7.38	1.0622	15.95	1.04758	8.6
$\hat{\alpha}_3$	1.14032	7.43	1.1634	16.14	1.13939	8.53
$\hat{\alpha}_4$	0.88626	5.73	0.91482	12.67	0.88859	6.56
$\hat{\alpha}_5$	0.85509	8.44	0.93369	16.33	0.86639	9.72
$\hat{\alpha}_6$	1.04669	7.65	1.0606	16.84	1.04204	8.84
$\hat{\alpha}_7$	0.97601	7.38	1.01853	15.6	0.98682	8.54
$\hat{\alpha}_8$	0.94781	6.65	0.98542	13.82	0.94771	7.6
$\hat{\alpha}_9$	0.95655	7.12	0.97719	15.4	0.95759	8.33
$\hat{\alpha}_{10}$	0.81215	8.9	0.9148	19.02	0.82722	10.36
$\hat{\alpha}_{11}$	0.90267	7.01	0.94522	14.93	0.91103	8.08
$\hat{\alpha}_{12}$	0.99056	5.65	1.01007	12.76	0.98673	6.45
$\hat{\alpha}_{13}$	1.03096	7.56	1.06153	15.63	1.0276	8.65
$\hat{\alpha}_{14}$	0.90194	8.95	1.00404	19.1	0.9238	10.42
$\hat{\alpha}_{15}$	1.04931	8.06	1.11424	17.55	1.05144	9.41
$\hat{\alpha}_{16}$	1.05438	7.11	1.0776	15.41	1.05206	8.19
$\hat{\alpha}_{17}$	1.10472	7.86	1.14869	17.12	1.10852	9.09
$\hat{\alpha}_{18}$	1.16242	8.45	1.21725	18.33	1.17266	9.87
$\hat{\alpha}_{19}$	1.22093	7.77	1.25507	16.99	1.21677	9
$\hat{\alpha}_{20}$	1.15979	8.99	1.29318	19.26	1.1764	10.46
$\hat{\alpha}_{21}$	1.37772	9.05	1.515	19.6	1.39444	10.54
$\hat{\alpha}_{22}$	1.55776	8.83	1.6395	19.44	1.55993	10.33
$\hat{\alpha}_{23}$	1.5496	8.65	1.65198	18.85	1.56343	10.1
$\hat{\alpha}_{24}$	1.53571	8.89	1.61513	19.33	1.53651	10.41
$\hat{\alpha}_{25}$	1.77596	8.98	1.85708	19.84	1.77433	10.5

Table 9. Continued

	1996–2009		1996–2003		1996–2008	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\alpha}_{26}$	1.98367	8.98	2.05421	19.84	1.96978	10.46
$\hat{\alpha}_{27}$	1.91648	8.94	1.99999	19.71	1.92248	10.46
$\hat{\alpha}_{28}$	1.97515	8.86	2.05081	19.54	1.96772	10.32
$\hat{\alpha}_{29}$	2.05311	8.96	2.14682	19.98	2.06648	10.48
$\hat{\alpha}_{30}$	1.96029	9.13	2.14069	20.15	1.98993	10.66
$\hat{\alpha}_{31}$	2.20211	9.04	2.30098	20.16	2.21198	10.57
$\hat{\alpha}_{32}$	2.19385	8.94	2.30038	19.86	2.20067	10.46
$\hat{\alpha}_{33}$	2.25753	9.07	-	-	2.25844	10.61
$\hat{\alpha}_{34}$	2.30651	9.11	-	-	2.30288	10.65
$\hat{\alpha}_{35}$	2.22761	9.06	-	-	2.24358	10.59
$\hat{\alpha}_{36}$	2.26332	9.01	-	-	2.28033	10.52
$\hat{\alpha}_{37}$	2.28565	9.09	-	-	2.2778	10.63
$\hat{\alpha}_{38}$	2.31307	9.14	-	-	2.32675	10.68
$\hat{\alpha}_{39}$	2.32874	9.08	-	-	2.33781	10.62
$\hat{\alpha}_{40}$	2.33858	8.94	-	-	2.3332	10.45
$\hat{\alpha}_{41}$	2.48442	9.11	-	-	2.46782	10.65
$\hat{\alpha}_{42}$	2.4825	9.14	-	-	2.47822	10.68
$\hat{\alpha}_{43}$	2.44861	9.13	-	-	2.42885	10.68
$\hat{\alpha}_{44}$	2.41217	9.08	-	-	2.41371	10.63
$\hat{\alpha}_{45}$	2.56122	9.14	-	-	2.54368	10.69
$\hat{\alpha}_{46}$	2.64082	9.16	-	-	2.64164	10.71
$\hat{\alpha}_{47}$	2.66495	9.14	-	-	2.66203	10.7
$\hat{\alpha}_{48}$	2.68254	9.12	-	-	2.68008	10.66
$\hat{\alpha}_{49}$	2.81145	9.14	-	-	2.80901	10.69
α_{50}	2.98276	9.17	-	-	2.96145	10.73
$\hat{\alpha}_{51}$	2.84972	9.15	-	-	2.82838	10.7
$\hat{\alpha}_{52}$	2.96522	9.14	-	-	2.956	10.69
$\hat{\alpha}_{53}$	2.85257	9.16	-	-	-	-
$\hat{\alpha}_{54}$	3.02278	9.17	-	-	-	-
$\hat{\alpha}_{55}$	2.96293	9.17	-	-	-	-
$\hat{\gamma}$	0.00821	13.71	0.00928	10.46	0.00815	13.69
$\hat{\omega}$	0.00958	40.04	0.01362	29.48	0.01014	38.25
$\hat{\vartheta}_2$	1.1668	109.17	0.90846	63.36	1.09962	106.62
$\hat{\vartheta}_3$	1.45254	80.2	1.19659	49.21	1.40271	74.95
$\hat{\vartheta}_4$	1.71222	72.32	1.49421	48.88	1.66785	66.63
$\hat{\sigma}_2$	0.55125	131.43	0.59487	79.96	0.54796	120.42
$\hat{\sigma}_3$	0.27964	86.2	0.30929	61.2	0.28722	80.68
$\hat{\sigma}_4$	0.18943	59.74	0.20634	46.35	0.18856	62.46
$\hat{\rho}$	0.51334	63.49	0.54572	33.66	0.52224	61.37

Table 10. Builder's Model estmate over selected time periods.

	1996–2009		1996–2003		1996–2008	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\theta}_1$	18.9504	5.14	6.54139	3.21	18.7768	5.55
$\hat{\theta}_2$	51.5232	6.01	42.3167	7.14	49.6144	6.87
$\hat{\theta}_3$	37.792	6.22	27.4228	11.44	36.5397	7.19
$\hat{\theta}_4$	39.7257	6.07	30.6971	10.47	39.5565	6.98
$\hat{\theta}_5$	22.9675	6.11	11.3205	9.2	22.2855	6.98
$\hat{\theta}_6$	35.2053	6.2	24.2672	11.5	34.2126	7.17
$\hat{\theta}_7$	36.4712	6.21	25.772	11.38	35.0501	7.2
$\hat{\theta}_8$	51.3282	6.22	37.5624	11.06	50.6889	7.2
$\hat{\theta}_9$	74.1072	6.25	54.2516	12.18	76.195	7.26
$\hat{\theta}_{10}$	65.5811	6.25	43.2627	11.97	61.4206	7.25
$\hat{\theta}_{11}$	77.1784	6.21	63.6382	10.51	75.6586	7.2
$\hat{\theta}_{12}$	29.2054	6.01	20.9272	9.91	27.5688	6.88
$\hat{\theta}_{13}$	38.9776	6.17	30.5877	11.08	38.6834	7.13
$\hat{\theta}_{14}$	60.2137	6.24	41.4114	11.87	59.2883	7.24
$\hat{\theta}_{15}$	96.2797	6.2	-	-	52.7786	6.61
$\hat{\theta}_{16}$	58.3052	6.2	42.518	10.83	56.2445	7.16
$\hat{\theta}_{17}$	37.3536	6.22	28.79	11.53	37.3925	7.19
$\hat{\theta}_{18}$	37.2665	6.14	23.0935	10.54	35.654	7.09
$\hat{\theta}_{19}$	45.4616	6.21	40.2212	11.53	47.4693	7.19
$\hat{\theta}_{20}$	38.1906	6.17	32.9879	11.05	54.0188	7.13
$\hat{\theta}_{21}$	44.821	5.89	30.7749	9.59	44.4373	6.68
$\hat{\theta}_{22}$	81.1283	6.26	59.7162	12.32	80.3557	7.27
$\hat{\alpha}_2$	0.95139	4.59	0.97994	9.69	0.95498	5.32
$\hat{\alpha}_3$	0.99103	4.78	1.03985	10.45	1.00201	5.57
$\hat{\alpha}_4$	0.7312	3.73	0.74296	7.74	0.75111	4.35
$\hat{\alpha}_5$	0.8116	5.01	0.87151	9.16	0.82589	5.87
$\hat{\alpha}_6$	0.95248	5	0.94333	10.8	0.95836	5.81
$\hat{\alpha}_7$	0.86708	4.99	0.88835	10.47	0.88542	5.78
$\hat{\alpha}_8$	0.84822	4.43	0.85083	8.73	0.86098	5.14
$\hat{\alpha}_9$	0.83986	4.65	0.82827	9.86	0.85809	5.44
$\hat{\alpha}_{10}$	0.82253	5.94	0.85865	12.16	0.83475	6.96
$\hat{\alpha}_{11}$	0.83732	4.27	0.82706	8.83	0.85551	4.93
$\hat{\alpha}_{12}$	0.90912	3.97	0.86057	7.86	0.92308	4.66
$\hat{\alpha}_{13}$	0.96641	4.94	0.93825	9.97	0.97604	5.68
$\hat{\alpha}_{14}$	0.8793	5.94	0.90544	11.97	0.89432	6.95
$\hat{\alpha}_{15}$	1.08107	5.38	1.06596	11.46	1.08874	6.3
$\hat{\alpha}_{16}$	1.06128	4.85	1.03543	10.05	1.06836	5.7
$\hat{\alpha}_{17}$	1.17468	5.72	1.13803	12.11	1.18416	6.66
$\hat{\alpha}_{18}$	1.23711	5.87	1.19014	12.49	1.25593	6.83
$\hat{\alpha}_{19}$	1.38538	5.45	1.32108	11.49	1.38288	6.32
$\hat{\alpha}_{20}$	1.35506	6.15	1.43629	12.9	1.36738	7.17
$\hat{\alpha}_{21}$	1.82033	6.2	1.86218	13.25	1.80975	7.23
$\hat{\alpha}_{22}$	2.00595	6.04	1.98121	13.1	1.99718	7.02
$\hat{\alpha}_{23}$	2.0415	6.02	2.06597	13.02	2.0506	7.01
$\hat{\alpha}_{24}$	2.052	6.1	2.01282	13.11	2.04297	7.1
$\hat{\alpha}_{25}$	2.38189	6.22	2.34719	13.58	2.36248	7.24
$\hat{\alpha}_{26}$	2.76026	6.21	2.72731	13.59	2.71551	7.23
$\hat{\alpha}_{27}$	2.75262	6.2	2.71339	13.55	2.73398	7.22

Table 10. Continued

	1996–2009		1996–2003		1996–2008	
Coefficient	Estimate	T stat	Estimate	T stat	Estimate	T stat
$\hat{\alpha}_{28}$	2.77131	6.13	2.77153	13.38	2.74748	7.13
$\hat{\alpha}_{29}$	2.92292	6.2	2.90953	13.62	2.91331	7.22
$\hat{\alpha}_{30}$	2.93856	6.28	2.9939	13.69	2.91188	7.31
$\hat{\alpha}_{31}$	3.23666	6.24	3.22505	13.65	3.21179	7.27
$\hat{\alpha}_{32}$	3.26326	6.21	3.28622	13.59	3.23435	7.23
$\hat{\alpha}_{33}$	3.35212	6.27	-	-	3.31537	7.3
$\hat{\alpha}_{34}$	3.38355	6.28	-	-	3.3352	7.32
$\hat{\alpha}_{35}$	3.2087	6.26	-	-	3.20256	7.3
$\hat{\alpha}_{36}$	3.27093	6.23	-	-	3.27335	7.25
$\hat{\alpha}_{37}$	3.29137	6.27	-	-	3.26951	7.31
$\hat{\alpha}_{38}$	3.37566	6.28	-	-	3.37289	7.32
$\hat{\alpha}_{39}$	3.31833	6.26	-	-	3.30773	7.3
$\hat{\alpha}_{40}$	3.34384	6.21	-	-	3.30449	7.24
$\hat{\alpha}_{41}$	3.63297	6.27	-	-	3.58544	7.31
$\hat{\alpha}_{42}$	3.48696	6.28	-	-	3.46523	7.33
$\hat{\alpha}_{43}$	3.51955	6.28	-	-	3.49224	7.32
$\hat{\alpha}_{44}$	3.3684	6.25	-	-	3.35519	7.29
$\hat{\alpha}_{45}$	3.65136	6.28	-	-	3.61947	7.32
$\hat{\alpha}_{46}$	3.76422	6.29	-	-	3.73952	7.33
$\hat{\alpha}_{47}$	3.7964	6.28	-	-	3.76926	7.33
$\hat{\alpha}_{48}$	3.87238	6.27	-	-	3.85314	7.31
$\hat{\alpha}_{49}$	3.98489	6.29	-	-	3.96809	7.33
α_{50}	4.20267	6.29	-	-	4.16664	7.34
$\hat{\alpha}_{51}$	4.12712	6.28	-	-	4.08914	7.32
$\hat{\alpha}_{52}$	4.18693	6.28	-	-	4.10956	7.32
$\hat{\alpha}_{53}$	4.1375	6.29	-	-	-	-
$\hat{\alpha}_{54}$	4.3602	6.3	-	-	-	-
$\hat{\alpha}_{55}$	4.4187	6.3	-	-	-	-
$\hat{\gamma}$	0.01035	14.87	0.01271	10.81	0.01	14.45
$\hat{\omega}$	0.01202	37.52	0.01869	26.34	0.01233	35.84
$\hat{\vartheta}_2$	1.11502	95.13	0.85087	49.33	1.06155	91.11
$\hat{\vartheta}_3$	1.42347	72.63	1.12454	41.79	1.41089	70.18
$\hat{\vartheta}_4$	1.58045	64.35	1.3149	40.44	1.59713	62.24
$\hat{\sigma}_2$	0.56898	120.51	0.60896	64.85	0.59141	109.21
$\hat{\sigma}_3$	0.30705	72.27	0.34388	45.43	0.32543	67.53
$\hat{\sigma}_4$	0.20825	55.63	0.23093	38.31	0.21331	56.84
$\hat{\rho}$	0.10284	6.12	-0.09809	-2.01	0.08523	4.76
$\hat{\delta}$	0.02351	46.04	0.01281	15.88	0.02494	46.12
$\hat{\tau}_2$	1.5368	78.92	1.44034	58.67	1.60901	80.03
$\hat{\tau}_3$	1.5062	57.6	1.3037	43.2	1.55714	55.62
$\hat{\varphi}_2$	1.38799	89.44	1.37268	80.35	1.36278	91.21
$\hat{\varphi}_3$	1.54054	30.65	1.58115	35.99	1.55477	33.45
$\hat{\varphi}_2$	1.24776	137.27	1.18751	100.68	1.21771	134.86
$\hat{\eta}_2$	1.23058	133.3	1.0808	90.13	1.20633	126.45

9. References

Davis, M., and M. Palumbo. 2008. “The Price of Residential Land in Large US Cities.” *Journal of Urban Economics* 63: 352–384. DOI: https://doi.org/10.1016/j.jue.2007.02.003.

De Haan, J., and W.E. Diewert. 2013. “Hedonic Regression Methods.” In *Handbook on Residential Property Prices Indices*, edited by J. De Hann and W.E. Diewert. (pp. 49–64). Luxembourg: Publication Office of the European Union, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF (accessed November 2019).

Diewert, W.E. 2009. “Basic Index Number Theory.” In *Consumer Price Index Manual: Theory and Practice*, edited by P. Hill. (pp. 263–268). Geneva: International Labour Organization (ILO). Available at: https://www.ilo.org/wcmsp5/groups/public/--dgreports/--stat/documents/presentation/wcms_331153.pdf (accessed November 2019).

Diewert, W.E. 2013. “Decomposing an RPPI into Land and Structure Components.” In *Handbook on Residential Property Prices Indices*, edited by J. De Hann and W.E. Diewert. (pp. 82–99). Luxembourg: Publication Office of the European Union, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF (accessed November 2019).

Diewert, W.E., and C. Shimizu. 2016. “Hedonic Regression Models for Tokyo Condominium Sales.” *Regional Science and Urban Economics* 60: 300–315. DOI: https://doi.org/10.1016/j.regsciurbeco.2016.08.002.

Diewert, W.E., and C. Shimizu. 2017. “Alternative Approaches to Commercial Property Price Indexes for Tokyo.” *Review of Income and Wealth* 63(3): 492–519. DOI: http://doi.org/10.1111/roiw.12229.

Prasad, N., and A. Richards. 2006. “Measuring Housing Price Growth: Using Stratification to Improve Median-Based Measures.” In *Research Discussion Paper* 2006-04, Reserve Bank of Australia. Available at: https://www.rba.gov.au/publications/rdp/2006/pdf/rdp2006-04.pdf (accessed November 2019)

Read-Hobman, T. 2015. “Evolution of Housing in Canada, 1957 to 2014.” In *Canadian Megatrends*. Ottawa: Statistics Canada, Catalogue no. 11-630-X. Available at: https://www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2015007-eng.pdf (accessed November 2019).

Received November 2019
Revised July 2020
Accepted August 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 803–825, http://dx.doi.org/10.2478/JOS-2020-0039

An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets

Marco Fortini[1]

Record linkage addresses the problem of identifying pairs of records coming from different sources and referred to the same unit of interest. Fellegi and Sunter propose an optimal statistical test in order to assign the match status to the candidate pairs, in which the needed parameters are obtained through EM algorithm directly applied to the set of candidate pairs, without recourse to training data. However, this procedure has a quadratic complexity as the two lists to be matched grow. In addition, a large bias of EM-estimated parameters is also produced in this case, so that the problem is tackled by reducing the set of candidate pairs through filtering methods such as blocking. Unfortunately, the probability that excluded pairs would be actually true-matches cannot be assessed through such methods.

The present work proposes an efficient approach in which the comparison of records between lists are minimised while the EM estimates are modified by modelling tables with structural zeros in order to obtain unbiased estimates of the parameters. Improvement achieved by the suggested method is shown by means of simulations and an application based on real data.

Key words: Structural zeros; robustness; EM algorithm; blocking.

1. Introduction

Record linkage (RL) consists of identifying pairs of records concerning the same individual, henceforth called matches, when these are included in different files or are duplicates in the same file. Matching is established by comparing a group of variables (key variables) that are common to the records and combined into a unique identifier. When the identifier is perfectly accurate the problem can be solved quite easily, and gains statistical interest if the key variables are affected by measurement errors.

Newcombe et al. (1959) laid the foundations for the probabilistic setting of the record linkage, while Fellegi and Sunter (1969) defined an optimal test for the identification of the matches amongst all the pairs of the Cartesian product between the files to be linked. The Fellegi-Sunter theory remains at the core of most of the current applications. In their approach, the statistical parameters of the test are estimated by the method of moments applied to the frequency distribution of the pairs from the Cartesian product by pattern of agreement between key variables. In so doing, the estimates are obtained without referring to a training set in which the match status of the pairs is known in advance. Winkler (1988) and Jaro (1989) improved the estimation procedure by introducing latent class models

[1] Italian National Institute of Statistics (Istat) – Directorate for Methodology and Statistical Process Design, Via C. Balbo, 16, 00184, Rome, Italy. Email: fortini@istat.it

(LCM), in which dichotomous indicators of agreement amongst key variables are used to identify the two-class latent variable that constitute the matching status of the pairs through the Expectation-Maximisation (EM) algorithm (Dempster et al. 1977).

As the size of the files to be linked grow, the size of their Cartesian product increases in a quadratic way, while the number of matches remains roughly proportional to the smaller of the two files. In these circumstances, the parameters of the LCM become biased quite early due to the disproportion between matches and non-matches and, thus, become useless to the purpose of record linkage.

This drawback in the estimation process is found well before the size of the pairs to be examined reaches the limits of the computational power available nowadays. As a matter of fact, it already occurs when the proportion p of matches out of the pairs of the Cartesian product falls below the limit of 5% (Winkler 2006) and key variables are not highly informative.

Empirical studies with the record linkage toolkit RELAIS (Cibella et al. 2009) developed by Italian National Institute of Statistics (Istat) show that using highly discriminant key variables, the above mentioned proportion p can be reduced up to 0.5–0.1% before it leads to the misclassification of matched and non-matched pairs and to the instability of the estimates. In other words, if we consider two files of the same size, a proportion p of 5% would limit file sizes to only $0.05^{-1} = 20$ units (two files of 20 records that produce a Cartesian product of 400 pairs), that could increase until 1,000 records (with a Cartesian product of a million pairs) in the best-case scenario, where estimates remain unbiased until a match rate of 0.1%. In order to go beyond this limit, Yancey (2002) proposed to enrich the density of matches in the set of checked pairs to obtain improved EM algorithm estimates for the record linkage conditional probability parameters.

In real applications, files of hundreds of thousands or even millions of records often have to be linked together, resulting in Cartesian product of thousands of billions of pairs and making the problem complicated quite early also in terms of computational resources. For these reasons, filtering (or indexing) methods have been largely proposed in literature to discard large numbers of the pairs with low evidence of being matches (Christen 2012 chap. 4) before the linkage estimation phase. Besides the great operational value achieved by these procedures (Baxter et al. 2003), filtering methods represent a step backwards from the probabilistic approach, since they condition the linkage probabilities to a subset of Cartesian product chosen with a deterministic criterion (Murray 2015). The adoption of this approach does not enable us to assess the risk of erroneous exclusion of real matches from the probabilistic linkage. In addition, it could cause false matches for those records whose true match pair was excluded from the analysis because of previous faults in the filtering process. Murray (2015) adapted estimation procedures for linkage parameter so as to take into account conditioning due to filtering process.

The contributions of this work are two-fold:

1. First of all, a sample filtering criterion of a subset of pairs that is representative of the entire Cartesian product of the two files to be combined is proposed. Unlike traditional filtering methods, a sampling approach makes it possible to evaluate matching probabilities for all of the pairs of the Cartesian product;

2. Secondly, the LCM parameters estimation are adapted through a robust EM approach which improves the standard one, as it remains unbiased when the proportion p of matches over the size of the Cartesian set becomes very small.

In this way, the number of pairs examined is reduced while the parameter estimation remains representative of the entire set of possible pairs. Although the robust estimation method and the sample-based filtering criterion are presented jointly in this article, it is worth noting that the former is also valid without the latter and can be used either with any other filtering criteria or along with the whole Cartesian set.

The remainder of the present article is organised as follows: Section 2 introduces the sample-based filtering approach after a formal definition of the record linkage and a description of the contingency table of pairs by their patterns of agreement between key variables, which is required for estimation purposes. Subsequently, in Section 3, the standard estimation method for record linkage parameters and its robust counterpart are shown. In Section 4, the accuracy of the present proposal is investigated by means of a simulation study, while Section 5 describes a real data example in which the method is implemented and is compared with a standard estimation approach based on traditional filtering. Lastly, in the final Section 6, some advantages and drawbacks are discussed, as well as potential further research to be carried out.

2. The Record Linkage Setting and the Dataset Under the Sample-Based Filtering Approach

In the first part of this section the canonical probabilistic record linkage is introduced along with a description of the data structure needed for the parameter estimation which will be described in Section 3.

As mentioned in the introduction, a simplistic matching of two files requires comparing each pair of records, which is inefficient and infeasible, particularly when files are large. As such, it is important to use some techniques to reduce the number of comparisons required. For our purposes, we use filtering, which first cycles through the record pairs and retains only those that match on at least a certain number of key variables. Filtering, then, assigns zero probability to records that do not match as they are considered extremely rare and discard them from the subsequent analyses. However, filtering does not allow to verify how actually close to zero is the probability of the excluded matches. Therefore, in the remainder of Section 2 a sample-based filtering approach is described as a way to make inferences on the complete Cartesian product when its size explodes due to the growth of the files to be merged.

Fellegi and Sunter (1969) consider record linkage as a discriminant analysis that, following their notation, aims to determine if a given pair of records (a, b), coming from the Cartesian product Ω between two files A and B to be linked, refers to the same unit. Let us denote as $\gamma = \{\gamma_1, . . ., \gamma_k\}$ the K values vector (or pattern) of agreement ($\gamma_i = 1$) or disagreement ($\gamma_i = 0$) between K couples of common key variables $(\mathrm{X}_{\mathrm{i},a}, \mathrm{X}_{\mathrm{i},b}; \mathrm{i} = 1, . . ., \mathrm{K}, (a, b) \in \Omega)$ measured on each pair of records $(a, b) \in \Omega$). We also denote Γ the set of 2^K patterns γ given by all the possible combinations of agreement/disagreement between the K key variables. Moreover, let us define as "matched" those pairs (a, b) whose records refer to the same unit and call M the corresponding set of matched pairs. Similarly, we define the set U as consisting of the pairs linking two different units and called “unmatched” pairs, with $\Omega = M \cup U$, and $M \cap U = \emptyset$.

The probabilities that the pair $(a, b) \in \Omega)$ has a given pattern of agreement γ given its match status are indicated as $m_\gamma = \mathrm{P}(\gamma|(a, b) \in M)$ for matched pair $(a, b) \in M$ and $u_\gamma = \mathrm{P}(\gamma|(a, b) \in U)$ for unmatched pair $(a, b) \in U$, while $p = P((a, b) \in M)$ represents the marginal proportion of matches $(a, b) \in M$ out of Ω.

Given the above notation, the likelihood ratio test $\mathrm{r}_\gamma = \frac{m_\gamma}{u_\gamma}$ is used by Fellegi and Sunter to assign the unknown matching status to pairs showing the vector γ. A pair is tagged as ‘matched’ when its corresponding r_γ is larger than a fixed threshold λ_M, and labelled instead as ‘unmatched’ when its rate r_γ is smaller than another fixed threshold $\lambda_U \leq \lambda_M$. When $\lambda_M > \lambda_U$ a certain number of pairs whose corresponding r_γ do not comply with none of the two limits, the pairs remain undecided and are sent to clerical review to be resolved. They showed that r_γ is the best among all the possible test statistics for the same problem and define a criterion to set λ_M and λ_U based on expected false positive and false negative error probabilities caused by wrong decisions, so as to obtain the right trade-off between the risk of error and the clerical review efforts.

It is important to note that pairs having the same pattern of agreement share the evidence of being matches, thus, all available information on the pairs’ status is enclosed in the frequency distribution of the pairs $(a, b) \in \Omega$ according to their γ pattern. In Section 3, we will see how probabilities m_γ, u_γ, $\forall\gamma$, and p – which were needed to accomplish the test – can be estimated from the frequency distribution of the pairs in Ω according to their patterns of agreement γ. Here we describe the characteristics of the frequency distribution of pairs, showed without loss of generality in Table 1 for three key variables, where 1 stands for agreement and 0 for disagreement. Subsequently, we will see how the table can be inferred under the sample-based filtering approach.

The first three columns of Table 1 describe the comparison vectors γ, with $\gamma_i \in (0, 1)$, $i = 1, 2, 3$. The fourth column represents the absolute frequencies of pairs N_γ within the pattern γ, with $\Sigma_\gamma N_\gamma = N_\Omega$, while the last column reports the relative frequencies. When K key variables are available, the frequency distribution is made up by 2^K frequency cells corresponding to all the possible combinations of agreement/disagreement patterns. When Ω increases due to the increasing size of the files, a sample-

Table 1. Contingency table of pairs by three key variables

γ_1	γ_2	γ_3	N_γ	p_γ
1	1	1	N_{111}	p_{111}
1	1	0	N_{110}	p_{110}
1	0	1	N_{101}	p_{101}
0	1	1	N_{011}	p_{011}
0	0	1	N_{001}	p_{001}
0	1	0	N_{010}	p_{010}
1	0	0	N_{100}	p_{100}
0	0	0	N_{000}	p_{000}
	Tot		N_Ω	1

based filtering of pairs from Ω is proposed in order to estimate the frequency distribution by the patterns γ, instead of investigating the whole set of pairs.

It is worth mentioning that in real cases, the comparison patterns separate matched and unmatched pairs quite well to the extent that unmatched pairs are expected to be found mainly within patterns with high disagreement. Similarly, matched pairs will fall mainly in patterns with little or no discrepancies. Since the set of unmatched pairs U is widely larger than the set M of matched pairs, the frequency distribution will concentrate toward the most discordant patterns. For this reason, a random sample would fail to accurately estimate the frequency cells for patterns showing large agreement. Consequently, our sampling approach is combined with the complete enumeration of pairs whose disagreement does not exceed one key variable. Full enumeration is obtained through K repeated merge tasks between the two files to be matched. This can be considered as a multiple blocking that uses all the sets of K-1 key variables adopted for parameters estimation on a rotating basis. Unlike standard multiple blocking, which recurs to variables (or coarser classifications) other than those used as keys at linkage stage, it is possible to do so because sampling from Cartesian product allows the shaping of comparisons table and parameters estimation, which would be impossible for conventional blocking.

Quite similarly to multiple blocking and other deterministic filtering techniques, this procedure allows for great computational savings in comparison to the processing of the whole set of pairs. Moreover, our approach outperforms the traditional filtering techniques as it allows for the evaluation of matching probabilities m_γ, u_γ, $\forall\gamma$ and p representative of the entire Ω through a robust EM approach, as will be explained in the next section.

The result of this procedure returns the quantities showed in Table 2 in the case of three comparison variables and is formally defined in the remainder of the present section for a generic number K of key variables.

More formally, in the sample-based filtering approach the frequencies N_γ of pairs $(a, b) \in \Omega$ by their pattern γ are entirely computed (or "solved" in what follows) only on the K+1 the patterns $\gamma \in \Gamma^S$, where at least all the K key variables but one agree; for this reason, Γ^S will be mentioned in the following as the set of solved patterns. Conversely, for residual patterns $\gamma \in \Gamma^R$, where the pairs disagree for more than one key variable, just the

Table 2. Estimated contingency table of the pairs by three key variables

γ_1	γ_2	γ_3	N_γ	p_γ
1	1	1	N_{111}	p_{111}
1	1	0	N_{110}	p_{110}
1	0	1	N_{101}	p_{101}
0	1	1	N_{011}	p_{011}
0	0	1	$\hat{N}_{001}$	$\hat{p}_{001}$
0	1	0	$\hat{N}_{010}$	$\hat{p}_{010}$
1	0	0	$\hat{N}_{100}$	$\hat{p}_{100}$
0	0	0	$\hat{N}_{000}$	$\hat{p}_{000}$
	Tot		N_Ω	1

overall number of pairs $N_{\Gamma^R} = \cup_{\gamma \in \Gamma^R} N_\gamma$ can be determined from

$$N_\Omega = \sum_{\gamma \in \Gamma^S} N_\gamma + N_{\Gamma^R}, \tag{1}$$

where $\Gamma^S \cup \Gamma^R = \Gamma$, $\Gamma^S \cup \Gamma^R = \varnothing$.

The frequencies N_γ, $\forall \gamma \in \Gamma^S$, can be operationally achieved through the following procedure:

1. execute K merge routines between the two files, one for each pattern γ that shows agreement on exactly K-1 variables, and find out the K amounts M_γ, $\forall \gamma \in \Gamma^S$, representing the number of pairs that are matched on each merge (many-to-many links are admitted),
2. calculate the number of pairs N_{γ^+} falling in the intersection of the merges made during the previous step 1, with N_{γ^+} representing the frequency for the pattern $\gamma^+ = \{\gamma_k = 1, \forall k \in (1, K)\}$ of agreement on all the K variables,
3. assign $N_\gamma = M_\gamma - N_{\gamma^+}$, for all that γ showing a pattern of agreement of order K-1,
4. being $N_\Omega = N_A N_B$ the product of records N_A, N_B in the two dataset to be linked, the residual frequency N_{Γ^R} is finally obtained from the formula (1).

Starting from a computational effort of $O(N^2)$ needed for comprehensive recognition of the pairs in Ω, in this way computations are reduced to the order of K times merge/sort routines, as the computational complexity of a merge/sort algorithm is $O(N \cdot log(N))$ (Cormen et al. 2009).

In addition, a simple random sample Ω^* of n pairs is selected from Ω, with distribution n_γ for $\gamma \in \Gamma$. The sample is then combined with the frequencies N_γ ($\forall \gamma \in \Gamma^S$) obtained from points 1-4 above to estimate the relative frequencies p_γ ($\forall \gamma \in \Gamma^R$) via the proportion

$$\hat{p}_\gamma = \frac{(N_\Omega - N_{\Gamma^S})}{N_\Omega} \cdot \frac{n_\gamma}{\sum\limits_{\gamma \in \Gamma^R} n_\gamma}, \quad \forall \gamma \in \Gamma^R. \tag{1}$$

The estimation of the frequencies N_γ, $\forall \gamma \in \Gamma^R$, is obtained through $\hat{N}_\gamma = N_\Omega \hat{p}_\gamma$, so as to shape data as shown in Table 2 for 3 key variables without loss of generality.

3. The Robust Parameter Estimation Method

In this section a proposal for robust estimation is presented after a brief illustration of the standard procedure for estimating the probability of record linkage.

In the traditional setting the conditional probabilities $m_\gamma = P(\gamma | (a, b) \in M)$, $u_\gamma = P(\gamma | (a, b) \in U)$, as well as the marginal proportion of matches in Ω, $p = ((a, b) \in M)$, are estimated (Jaro 1989) through the EM algorithm on the basis of the observed data approach and conditional independence assumptions

$$m_\gamma = \prod_k m_k; \qquad u_\gamma = \prod_k u_k; \quad \forall \gamma \tag{2}$$

where

$$m_k = \Pr\left(\gamma_k = 1|\left(a,b\right) \in M\right), \quad u_k = \Pr\left(\gamma_k = 1|\left(a,b\right) \in U\right), \quad k = 1,\ldots,K.$$

It is important to note, here and hereafter, the difference between the marginal probabilities m_k and u_k, for $k = 1,\ldots,K$, referred to single key variables and the combined probabilities m_γ and u_γ, $\forall\gamma$, concerning patterns.

To such purpose, Jaro define the 'augmented' likelihood for the complete dataset in which the proportion of matches g_γ among N_γ is known for every γ

$$l\left(g_\gamma, N_\gamma|\boldsymbol{m},\boldsymbol{u},p\right) = p^{N_M}\left(\prod_\gamma m_\gamma^{g_\gamma N_\gamma}\right)\left(1-p\right)^{\left(N_\Omega - N_M\right)}\left(\prod_\gamma u_\gamma^{\left(1-g_\gamma\right)N_\gamma}\right). \tag{3}$$

Since, as a matter of fact, the g_γ proportions are unknown, the maximum of the likelihood in (3) cannot be directly estimated and the estimates are achieved through an EM approach. After the $2{\cdot}K+1$ starting values m_k^0, u_k^0, for $k = 1,\ldots,K$, and p^0 are assigned, the step E is carried out to compute the expected values

$$g_\gamma^0 = P^0\left(\left(a,b\right) \in M|\gamma\right) = \frac{m_\gamma^0 p^0}{m_\gamma^0 p^0 + u_\gamma^0\left(1-p^0\right)}, \quad \forall\gamma$$

under the conditional independence assumption (2).

Once the expected values are obtained for g_γ, they can be used to maximise the likelihood in (3) during the step M so as to update the parameters for $k = 1,\ldots,K$. It can be shown (Jaro 1989) that MLE parameters are given by

$$p^1 = \frac{\sum_\gamma g_\gamma^0 N_\gamma}{N_\Omega}, \qquad m_k^1 = \frac{\sum_{\gamma:\gamma_k=1} g_\gamma^0 N_\gamma}{p^0 N_\Omega}, \qquad u_k^1 = \frac{\sum_{\gamma:\gamma_k=1}\left(1-g_\gamma^0\right)N_\gamma}{\left(1-p^0\right)N_\Omega}$$

when conditional independence (2) is assumed. The procedure, which is iterated until convergence is achieved, is proven to be stable and fairly insensitive to the starting values under conditional independence assumptions (Jaro 1989; Winkler 1988).

The robust procedure, described in what follows, deals with the bias affecting the standard method when Ω becomes larger and the pairs (a,b) ∈ M become very small in comparison with those in the whole Ω. The robust EM procedure uses multinomial models with incomplete tables (Bishop et al. 1975) for estimation of m_k and u_k parameters, $k = 1,\ldots,K$, during the **M** step of the EM. Step **M** is hereinafter mentioned in bold in order to be distinguished from the set *M* of the matched pairs.

As usual, the procedure begins with starting values m_k^0 and u_k^0, $k = 1,\ldots,K$, and p^0, so that the E step is carried out by computing the expected values

$$g_\gamma^0 = P^0\left(\left(a,b\right) \in M|\gamma\right) = \frac{m_\gamma^0 p^0}{m_\gamma^0 p^0 + u_\gamma^0\left(1-p^0\right)}, \quad \forall\gamma \in \Gamma.$$

Subsequently, the **M** step updates the probability p by using N_γ, $\gamma \in \Gamma^S$, by means of solved patterns along with estimated frequencies $\hat{N}_\gamma$ for $\gamma \in \Gamma^R$ from residual patterns

$$p^1 = \frac{\sum_{\gamma \in \Gamma^S} \left(g_\gamma^0 N_\gamma \right) + \sum_{\gamma \in \Gamma^R} \left(g_\gamma^0 \hat{N}_\gamma \right)}{N_\Omega}.$$

Now the **M** step for m_k and u_k estimation is modified to prevent the bias due to the disproportion between the sets M and U. It should be observed that, given the expected proportions g_γ^0 already determined at step E, the parameters m_k and u_k can always be estimated independently of each other by their respective part of the likelihood function in (3). For this reason, a description of how to obtain the m_k estimates will be provided in what follows.

Instead of using all the frequencies as in the standard case, robust approach estimates the m_k, $k = 1,. . .,K$, only from the K+1 expected counts $g_\gamma^0 N_\gamma$ for $\gamma \in \Gamma^S$, from solved patterns while the remaining patterns $\gamma \in \Gamma^R$ are considered as structural zeros and their frequencies are trimmed from the maximisation step.

This approach resembles the one adopted by other scholars (see Neycov et al. 2007), in which trimmed estimation is used in a cluster analysis context during the **M** step of the EM procedure to avoid inconsistencies in parameters estimation due to the presence of outliers. In the present case, the underlying idea is that m_k's (and u_k's) parameters estimation during the **M** step is more robust after trimming those patterns γ whose cells, in their respective conditional distributions to the M and U sets, are probably zero under the true model.

In so doing, the relevant part of likelihood (3) with respect to m_k, $k = 1,. . .,K$, can be reparametrised as

$$l\left(g_\gamma^0 N_\gamma, \phi \right) \propto \prod_{\gamma \in \Gamma^S} \phi_\gamma^{\left(g_\gamma^0 N_\gamma \right)}$$

where $\phi_\gamma = \frac{m_\gamma}{m_S}$ are the parameters conditioned to the probability mass $m_S = \sum_{\gamma \in \Gamma^S} m_\gamma$ on solved patterns Γ^S.

Indicating with $M_S = \sum_{\gamma \in \Gamma^S} \left(g_\gamma^0 N_\gamma \right)$ the sum of matched pairs expected in solved patterns and given that $m_S + \sum_{\gamma \in \Gamma^R} m_\gamma = 1$ by definition, estimates of ϕ_γ are provided by $\hat{\phi}_\gamma = g_\gamma^0 N_\gamma / M_S$, $\forall \gamma \in \Gamma^S$.

Given the relationship between the sets of parameters ϕ_γ and m_γ and the independence assumptions (2) that relate m_γ to their components $m_k, k = 1,. . .,K$, it can be written

$$\frac{\hat{\phi}_{\gamma^+}}{\hat{\phi}_{\gamma(-k)}} = \frac{m_k \prod_{h \neq k} m_h}{(1 - m_k) \prod_{h \neq k} m_h} = \frac{m_k}{(1 - m_k)} = \frac{g_{\gamma^+}^0 \cdot N_{\gamma^+}}{g_{\gamma(-k)}^0 \cdot N_{\gamma(-k)}}, \qquad k = 1,. . ., K$$

where:

- $\gamma^+ = \{\gamma_k = 1, \mathrm{k} = 1,. . ., \mathrm{K}\}$ is the pattern in which all the key variables agree;
- $\gamma(-\mathrm{k}) = \left\{ \gamma_k = 0, \gamma_i = 1, \mathrm{i} \in (1, \mathrm{K}), \forall \mathrm{i} \neq \mathrm{k} \right\}$ are the pattern in which only the k-th key variable disagrees (e.g., for K=3, $\gamma(-2) = \{1, 0, 1\}$).

Finally, with some algebraic computations the estimates of m_k, $k = 1, \ldots, K$ are achieved as

$$m_k^1 = \frac{g_{\gamma^+}^0 \cdot N_{\gamma^+}}{g_{\gamma^+}^0 \cdot N_{\gamma^+} + g_{\gamma(-k)}^0 \cdot N_{\gamma(-k)}}, \quad k = 1, \ldots, K, \tag{4}$$

Similarly, the u_k's are estimated through the multinomial distribution conditioned to the frequencies $\left(1 - g_\gamma^0\right) N_\gamma$ on the K+1 patterns γ such that $\sum_k \gamma_k \leq 1$ (i.e., patterns in which no more than one key variable agrees) and considering all the other patterns as structural zeros

$$u_k^1 = 1 - \frac{\left(1 - g_{\gamma^-}^0\right) N_{\gamma^-}}{\left(1 - g_{\gamma^-}^0\right) N_{\gamma^-} + \left(1 - g_{\gamma(k)}^0\right) N_{\gamma(k)}}, \quad k = 1, \ldots, K \tag{5}$$

where

- $\gamma^- = \{\gamma_k = 0, \ k = 1, \ldots, K\}$ is the pattern in which all key variables disagree;
- $\gamma(k) = \left\{\gamma_k = 1, \gamma_i = 0, \ i \in \left(1, K\right), \forall \, i \neq k\right\}$ are the patterns in which only the k-th key variable agrees (e.g., for K=3, $\gamma(2) = \{0, 1, 0\}$).

Again, the procedure is iterated until convergence is reached.

The probabilities $g_\gamma = P\left(\left(a, b\right) \in M | \gamma\right)$ computed during the E step of EM procedure play another important role in the strategy of assigning pairs to the sets U and M. Larsen and Rubin (2001) showed that g_γ is a monotonic transformation of r_γ and therefore, inducing the same order on the pairs, these can be used interchangeably. Moreover, g_γ has a more immediate interpretation and can also be used to estimate the number of pairs incorrectly assigned either to M or U sets. In fact, summing up the corresponding g_γ for all the pairs assigned to U it is possible to obtain the expected number of missed matches. Conversely, by adding the corresponding $1 - g_\gamma$ for all the pairs assigned to M, the expected number of false matches included among linked units is obtained. These concepts have been taken into account as they will be used in Sections 4 and 5 to assess the proposed procedure.

It is worth noting that robust estimation approach can be applied regardless of the sampling procedure shown in Section 2. In fact, it may be also used when 'lighter' filtering is preferred in order to keep the risk of exclusion of true matches from analysis as low as possible but the ratio between matches and non-matches does not make it possible for the standard estimation method to work.

In the following section, the behaviour of the present approach will be investigated in a simulation context so as to prove its ability in attaining accurate estimates when the match rate falls even far below 0.1%, limit under which the traditional approach is known as not available.

4. Simulation Study

As show above, the main limit of classic filtering is that it does not allow the estimation of the probability of matching for all possible pairs from Cartesian product between the files to be matched. Consequently, it is not possible to guess the number of unmatched pairs that could be caused by filtering errors. Although the present method allows overcoming this limit, it is up to us to show that the accuracy of the estimates obtained is adequate.

As a matter of fact, when considering the present proposal, the accuracy of the estimates can be affected by the following conditions:

1. The sample size of Ω^* from Ω, since the estimates $\hat{p}_\gamma$ of the distribution $p_\gamma, \forall \gamma \in \Gamma$, are affected by sampling error,
2. The number K of key variables, because the possible patterns increases as a function of 2^K, while robust estimation approach uses only $2 \cdot (K + 1)$ of them, that is, those patterns with either the highest agreement or disagreement between the key variables respectively for M and U sets,
3. The magnitude of $u_k, k = 1, \ldots, K$ probabilities, since even a small increase of u_k can raise the number of unmatched pairs that agree on many key variables, when applied to the big amount of pairs $(a,b) \in U$.

This section investigates the accuracy of estimates with respect to the previous points by carrying out a simulation. In order to test the accuracy, various true contingency tables for pairs $(a,b) \in \Omega$ according to patterns of agreement $\gamma \in \Gamma$ between the K key variables are generated by specifying the following parameters:

- size N_A and N_B of the files A and B to be linked,
- size N_M of the match set M,
- number K of key variables,
- sampling rate f of pairs selected from Ω,
- number of false matches ε_0, defined in this analysis as the number of pairs $(a,b) \in U$ whose pattern of agreement is $\gamma \in \Gamma^S$, and
- number of missed matches ε_1, defined in this analysis as the number of pairs $(a,b) \in M$ whose pattern of agreement is $\gamma \notin \Gamma^S$.

For given values of K, ε_0 and ε_1, the m_k and $u_k, k = 1, \ldots, K$, probabilities are obtained from equations determining the sum of expected frequencies in patterns $\gamma \in \Gamma^S$ for non-matched and matched pairs, respectively,

$$(N_A N_B - N_M) P\big(\gamma \in \Gamma^S | U\big) = \big(N^2 - N_M\big)\, \big(u^K + k \cdot u^{K-1}(1-u)\big) = \varepsilon_0 \qquad (6)$$

$$N_M\big(1 - P\big(\gamma \in \Gamma^S | M\big)\big) = N_M\big(1 - \big(m^K + k \cdot m^{K-1}(1-m)\big)\big) = \varepsilon_1 \qquad (7)$$

where conditional independence (3) is assumed and further restrictions $u_k = u$ and $m_k = m$, for $k = 1, \ldots, K$, are adopted for simplicity.

Once the true frequency table $N_\gamma, \forall \gamma \in \Gamma$ is obtained through parameters n, m and $p = \frac{N_M}{N_A N_B}$, the observed frequencies for solved patterns $N_\gamma, \forall \gamma \in \Gamma^S$ follow directly as a sub-table. Sampling frequencies for residual patterns n_γ, $\gamma \in \Gamma^{\mathrm{R}}$ are instead obtained by sampling from the conditional multinomial distribution with parameters $p_\gamma = N_\gamma / N_{\Gamma^{\mathrm{R}}}, \forall \gamma \in \Gamma^{\mathrm{R}}$ with sampling ratio f and estimates of $N_\gamma, \forall \gamma \in \Gamma^{\mathrm{R}}$ finally resulting from

$$\hat{N}_\gamma = \frac{n_\gamma \left(N_A N_B - \sum_{\gamma \in \Gamma^S} N_\gamma \right)}{\sum_{\gamma \in \Gamma^R} n_\gamma}, \forall \gamma \in \Gamma^{\mathrm{R}}.$$

In the next simulation, 400 sampling tables n_γ, $\boldsymbol{\gamma} \in \Gamma^R$ were generated by Monte Carlo trials for every set of parameter combinations. A file size of one million records each were considered for files A and B, assuming a perfect match between them, that results in a set M of one million matches, a set U of 999.999 billion of non-matches and a marginal matching probability p = 1e-6, far below the limit that undermines the standard estimation process. These three parameters were kept fixed during the analysis for the sake of brevity, as they did not show any influence on accuracy during non-systematic checks carried out in advance on the data. For every sample, the robust EM algorithm was applied in order to estimate linkage parameters m_k, u_k, $k = 1, \ldots, K$ and p and estimate of $\hat{\varepsilon}_0$ and $\hat{\varepsilon}_1$ are obtained through Equations (6) and (7). Starting values for EM procedure were kept fixed to $m_k = 0.9$, $u_k = 0.1$ $k = 1, \ldots, K$ and $p = 0.01$ after having verified that they do not affect the convergence of the likelihood to its global optimum.

The first evaluation concerned the influence of sampling rate on the accuracy of estimated amounts of false and missed matches. Five sampling rates *f* were tested, (0.3, 0.6, 1.0, 1.4, 1.7 per million), resulting in samples varying from 300,000 to 1.7 million of pairs.

Figures 1a and 1b below report the sampling distribution for the estimate $\hat{\varepsilon}_1$ of missed matches according to increasing sampling rates, for five and nine key variables, when *u* is fixed so as to provide $\varepsilon_0 = 5{,}000$ false matches by Equation (6). Both for the present and the following diagrams, the solid line represents the true value, the dashed line describes the sampling median, the dot-dashed line indicates the sampling mean and dotted lines identify respectively 5% and 95% quantiles of the sampling distribution. It is possible to note that the distribution of missed matches $\hat{\varepsilon}_1$ tends to be less variable and fairly symmetrical for K=5. The distribution is more skewed for *K*=9, and its main part is placed slightly under the true value, since the median and the 5% quantile both lie under the true value and are very close one another. However, some estimates can be upward biased, as proven by the average value overlaying the true one, while the 95% quantile is much larger.

Figures 2a and 2b show the sampling distribution of false matches estimates $\hat{\varepsilon}_0$ according to increasing sampling rates, for five and nine key variables, when *m* is fixed so as to provide $\varepsilon_1 = 10{,}000$ missed matches by Equation (7). Here, it can be noted that

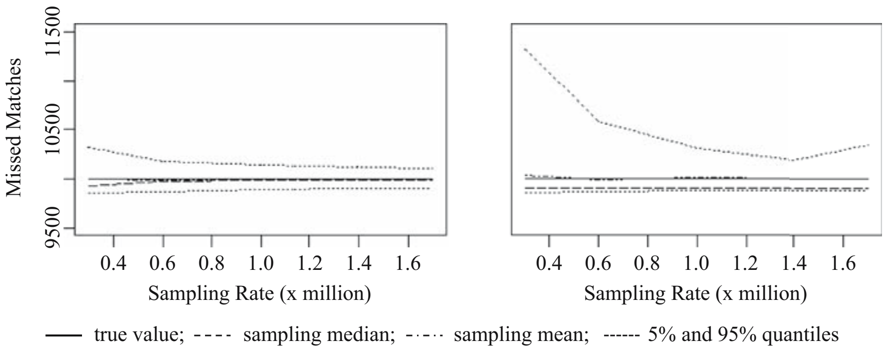


Fig. 1. a. Estimate of 10,000 missed matches by sampling rate – five key variables. b. Estimate of 10,000 missed matches by sampling rate – nine key variables

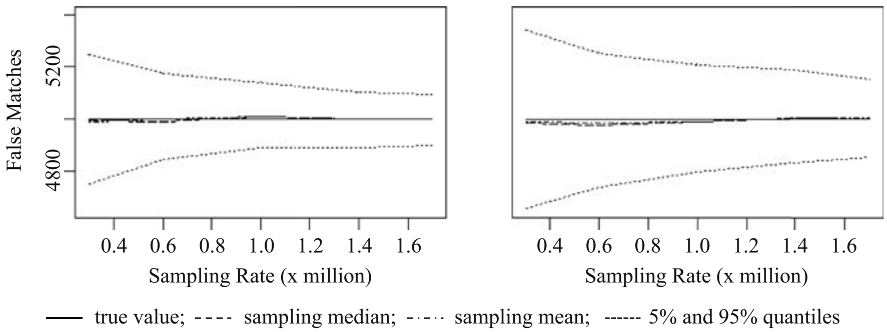


Fig. 2. a. Estimate of 5,000 false matches by sampling rate – five key variables. b. Estimate of 5,000 false matches by sampling rate – nine key variables.

distribution of $\hat{\varepsilon}_0$, being mostly of little variability and quite symmetrical for practical purposes, shows less variability for K=5 and for large sampling rates.

Figure 3 summarises the sample distribution of the estimated missed matches amount $\hat{\varepsilon}_1$ according to increasing number of false matches, in case of five key variables. A sample of one million pairs was considered for these tests. Surprisingly enough, an anomalous behaviour of the sampling distribution is observed for missed matches estimates when the incidence of false matches is very small. The median of the distribution is very close to the 5% quantile and both are very close to the true value.

The 95% quantile is instead much larger than the true value, denoting the risk of outliers in the estimate of ε_1 when data are affected by only few false matches (i.e., probabilities $u_{k,}\, k = 1,. . .,K$ are very small). When the number of false matches increases to 50,000 pairs, the sampling distribution of missed matches becomes symmetrical while its variability decreases, until the number of false matches becomes larger and variability starts to increase significantly again. It is worth noting that, even in the worst-case scenario, estimates are not further from the true value than about 10%.

In Figure 4, the distribution of missed matches $\hat{\varepsilon}_1$ as a function of false matches is reported for K=9. The distribution performs better than the corresponding one for K=5, since it is almost perfectly close to the real value when false matches are few, with negligible variability and a relative bias of less than 1%. The variability of the distribution

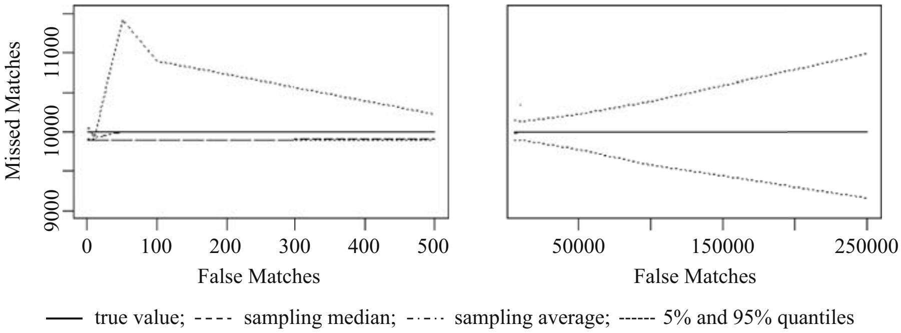


Fig. 3. Estimate of 10,000 missed matches by false matches – five key variables.

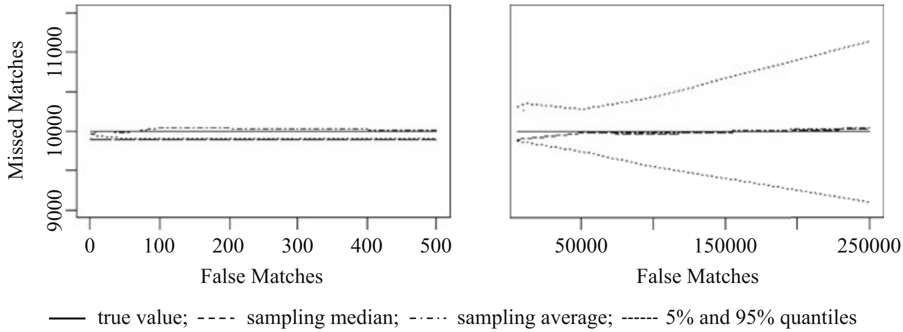


Fig. 4. Estimate of 10,000 missed matches by false matches – nine key variables.

continues to be small until the missed matches do not exceed 50,000, a quite pessimistic scenario corresponding to 5% of the total number of matches considered in the current simulation. This behaviour corroborates the accuracy of estimates when the number of key variables increases and at least one million pairs are sampled from the 10^{12} pairs in Ω.

Finally, Figures 5 and 6 show the sampling distribution of the estimated number of false matches $\hat{\varepsilon}_0$ by increasing the number of missed matches, for five and nine key variables,

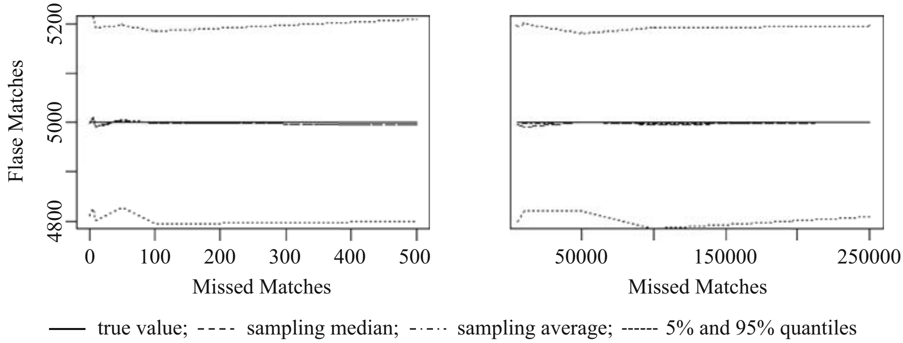


Fig. 5. Estimate of 5,000 false matches by missed matches – five key variables.

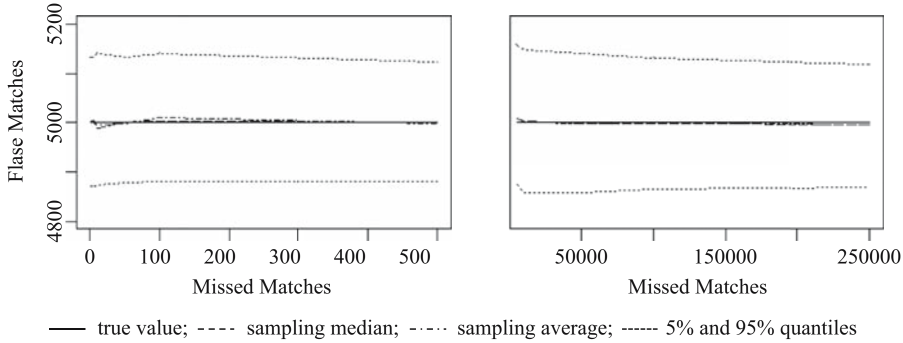


Fig. 6. Estimate of 5,000 false matches by missed matches – nine key variables.

respectively. It can be noted that the estimates are unbiased, have a symmetric distribution and are not affected by the incidence of missed matches, while the number of the key variables influences only marginally its variability.

5. A Real Case Study

This section provides an application of the method to real data and makes a comparison with a common filtering approach. The goal is to show how the linkage strategy presented above works when applied to a common case where (at least) one of the key variables is compared through a fuzzy string similarity distance (Herzog et al. 2007). Furthermore, a focus will be put on how the estimate of linkage probability for all the possible pairs can be used in practice for identifying further matches in patterns previously excluded by the filtering step.

The aim is not to make an exhaustive comparison between the present method and the most advanced filtering techniques. Such a comparison is beyond the scope of the present work since, to our knowledge, no advanced filtering method can return the linkage probability for pairs excluded from analysis, which our approach is able to do. For this reason, we choose to compare our approach with a sorted neighbour filtering (Hernandez and Stolfo 1995) which allows to appreciate the differences between approaches without unnecessarily complicating the exposition.

In our example, a file of 16,723 foreigners who applied for a permit to stay (PS) is linked to another file of 19,398 foreigners registered in the municipal population registry lists (MPR). Although the size of the two files is not excessively large, in this case the Cartesian product between them results in more than 324 million of possible pairs and the standard Fellegi-Sunter approach already needs of a filtering step in order to be applied without bias. Since, by definition, no more than one record of the larger file can be matched to each record of the smaller one, the proportion p of matches over all the possible pairs cannot exceed the rate of about $5 \cdot 10^{-5}$, in this case (1/19,398). The example is taken from a preparatory application to the 2011 Italian census, which aimed to fetch a contact list of foreigners not already registered in population registers among the appliers for a “permit to stay” (Fortini et al. 2013). The linkage meant to identify people in the PS file who were not already included in the population register, so as to be contacted during the census. In our example, data concerning the Abruzzo region are investigated in order to identify people enlisted in both sources.

The key variables used to identify individuals in both the lists are: First and last name (single field), Gender, Country of citizenship code, Day of birth, Month of birth and Year of birth.

Although robust estimation assisted by sampling filtering is, in our opinion, intrinsically better than classic approach as virtually including all possible pairs in the evaluation of their probability of linkage, the results of the proposed application will be briefly compared with those obtained by filtering with ’sorted neighbours’ algorithm on the variable ’First and last name’.

As far as the quality of input data is concerned, the amount of missing values affecting key variables can be considered negligible, since the MPR source is affected by missing data only on the variable ‘First and last name’ at a 0.01% level, while PS shows 5% of missing data on ’Gender’ and an abnormal concentration of births occurring on the first day of January (about 750 cases against average values between 50 and 70 for the other dates of the year), which could affect the discriminatory power of the day and month of birth.

The pairwise comparison between key variables is carried out by means of dichotomous indicators that assumes value 1 in case of agreement and 0 otherwise. The agreement is established when both the instances of the key variable have the same value, except for missing values whose presence always gives rise to a disagreement. Only for the variable ‘First and last names’, the comparison was grounded on the three-gram Jaccard string distance, which always ranges between 0 and 1 (Christen 2012), and is labelled as agreement when exceeding 0.7 and as disagreement otherwise.

In order to achieve the frequency distribution of pairs by solved patterns, the procedure presented in Section 2 was applied to all the key variables with exclusion of ‘First and last names’. In fact, the agreement/disagreement label to ‘First and last names’ was assigned after the evaluation of the Jaccard distance to all the pairs falling within the patterns recognised by the other key variables. With regard to the effort for completing this phase of data processing, 60,616 pairs were checked.

A random sample of 500,000 pairs was subsequently selected from the Cartesian product of the two files to be linked by means of two independent random selections with replacement of 500,000 records from the files and their consequent inclusion, which finally reduces to 499,642 pairs after the exclusion of duplicates. Table 3 shows the frequency distribution of pairs according to solved (bolded rows) and remaining patterns, whose frequencies are estimated through the sample. It can be noted that sampling zeros are achieved for some of the patterns characterized by high concordance between key variables, since their expected frequencies are fairly low. Since these patterns are not directly used for estimation of parameters m_k and u_k, these flaws do not invalidate the method. Nevertheless, they can affect the precision of the expected number of matches for these specific patterns and offer a prospect for future improvement on sample design and model estimation of observed frequencies for remaining patterns.

Data in Table 3 was used for estimation of linkage parameters with robust EM, as described in Section 3. Starting values $\mathrm{m_k} = 0.9$, $\mathrm{u_k} = 0.1$, $\mathrm{k} = 1,. . .,6$, were assigned to each of the six key variables, while starting value p was fixed at 0.01. Tests carried out with many different starting values does not change the estimates resulting from the method, but only affects the number of iterations needed for the algorithm to converge. Under these conditions, the robust EM algorithm converges in about 50 iterations, returning $\mathrm{p} = 1.86\mathrm{E} - 5$ and the m’s and u’s values shown in the Table 4.

From rates r_k between m_k and u_k, for $k = 1,. . .,6$, it can be seen that ‘First and last names’ has the highest discriminant power followed by ‘Day of birth’, ‘Year of birth’, ‘Country ID’, ‘Month of birth’ and finally ‘Gender’. Expected frequencies of matched and non-matched pairs by their patterns of agreement (solved patterns are reported in bold) are shown in Table 5, as well as the matching probabilities conditional to the pattern. The patterns are sorted by descending probabilities $P(M|\gamma)$ so as to identify those having a higher number of true matches with minimum inclusion of false ones.

By only considering the solved patterns and adopting a threshold of 0.5 on $P(M|\gamma)$, the number of matches expected to be retrieved is 5,172 plus one false match and 26 missed

Table 3. Contingency table of pairs by agreement patterns (solved patterns in bold).

Surnames Names	Gender	Country ID	Day of birth	Month of birth	Year of birth	Frequency	Surnames Names	Gender	Country ID	Day of birth	Month of birth	Year of birth	Frequency
1	**1**	**1**	**1**	**1**	**1**	**3691**	**0**	**1**	**1**	**1**	**1**	**1**	692
1	**1**	**1**	**1**	**1**	**0**	**34**	0	1	1	1	1	0	7711
1	**1**	**1**	**1**	**0**	**1**	**42**	0	1	1	1	0	1	2570
1	1	1	1	0	0	643	0	1	1	1	0	0	75178
1	**1**	**1**	**0**	**1**	**1**	**20**	0	1	1	0	1	1	17991
1	1	1	0	1	0	2570	0	1	1	0	1	0	622625
1	1	1	0	0	1	0	0	1	1	0	0	1	255732
1	1	1	0	0	0	21204	0	1	1	0	0	0	6970957
1	**1**	**0**	**1**	**1**	**1**	**529**	0	1	0	1	1	1	1285
1	1	0	1	1	0	0	0	1	0	1	1	0	131721
1	1	0	1	0	1	0	0	1	0	1	0	1	40480
1	1	0	1	0	0	643	0	1	0	1	0	0	1399460
1	1	0	0	1	1	0	0	1	0	0	1	1	389381
1	1	0	0	1	0	3213	0	1	0	0	1	0	11044684
1	1	0	0	0	1	643	0	1	0	0	0	1	4134126
1	1	0	0	0	0	22489	0	1	0	0	0	0	123255948
1	**0**	**1**	**1**	**1**	**1**	**195**	0	0	1	1	1	1	0
1	0	1	1	1	0	0	0	0	1	1	1	0	7068
1	0	1	1	0	1	0	0	0	1	1	0	1	3855
1	0	1	1	0	0	0	0	0	1	1	0	0	75820
1	0	1	0	1	1	0	0	0	1	0	1	1	22489
1	0	1	0	1	0	0	0	0	1	0	1	0	684952
1	0	1	0	0	1	643	0	0	1	0	0	1	226175

Table 3. Continued

Surnames Names	Gender	Country ID	Day of birth	Month of birth	Year of birth	Frequency	Surnames Names	Gender	Country ID	Day of birth	Month of birth	Year of birth	Frequency
1	0	1	0	0	0	9638	0	0	1	0	0	0	7444511
1	0	0	1	1	1	643	0	0	0	1	1	1	9638
1	0	0	1	1	0	0	0	0	0	1	1	0	159993
1	0	0	1	0	1	0	0	0	0	1	0	1	45621
1	0	0	1	0	0	0	0	0	0	1	0	0	1741936
1	0	0	0	1	1	0	0	0	0	0	1	1	431147
1	0	0	0	1	0	1285	0	0	0	0	1	0	12894567
1	0	0	0	0	1	0	0	0	0	0	0	1	4831929
1	0	0	0	0	0	10281	0	0	0	0	0	0	144040952

Table 4. Marginal probability of agreement for each key variable depending on the match status of the pair.

	First and last names	Gender	Country id	Day of birth	Month of birth	Year of birth
k	1	2	3	4	5	6
m_k	0.984	0.887	0.790	0.957	0.964	0.963
u_k	0.000	0.461	0.049	0.012	0.082	0.032

ones. In real cases, a certain amount of clerical work is spent to scrutinise pairs whose probability $P(M|\gamma)$ is lower than a fixed value so as to distinguish other true matches, which is not done here for practical reasons. However, if the whole table is considered, the number of expected matches increases to 6,037, with a difference that leads to examine other patterns that have high match probability $P(M|\gamma)$. Table 5 shows that 729 additional matches are expected from the scrutiny of nine residual patterns in which $P(M|\gamma) \geq 0.5$, together with a concomitant inclusion of 54 more false match pairs. Nonetheless, sampling variability affects this figures due to sample survey on the set of remaining patterns and, subsequent to investigation of nine patterns, the number of additional true matches is reduced to 273, with an inclusion of 29 more false matches.

This additional work requires the screening of 811,341 pairs which, added to the 60,616 and 500,000 pairs already considered, brings to 1,371,957 the total number of pairs taken into account, out of the more than 324 million of couples from the Cartesian product across the files. Consequently, 5,475 links are found overall, 30 of which are expected to be false, while 592 matches are expected to be missed across patterns with $P(M|\gamma) < 0.5$.

In order to compare the example above with the standard approach, the linkage exercise was repeated after filtering the set of pairs by a ‘Sorted neighbours’ procedure on the variable ‘First and last name’ with a comparison window of 50 units. Among other things, the attempt to expand the window to 100 units produces about 1.7 million pairs to analyse and leads to the failure of the standard estimation procedures of the probability of linkage. The adopted filtering selection sets the pairs to 861,528, on which the standard EM estimation approach works furtherly by returning a marginal probability $p = 0.006$. This probability is 200 times larger than the one estimated by the present method, supporting the fact that filtering increased the ratio between matches and non-matches among pairs retained in analysis. Vectors of parameters m_k and u_k, estimated through the standard method, are shown in Table 6 to be compared with those previously reported in Table 4. It can be noted that, despite a fairly similar overall values for m_k and u_k, the probabilities u_k related to non-match pairs increases for ’Country ID’ and ‘First and last Names’, with a change in their rank in terms of discriminatory power compared with other key variables.

On considering patterns that have posterior probabilities $\mathrm{P}(M|\boldsymbol{\gamma}) \geq 0.5$, the filtered approach gives rise to 5,129 links, 5,102 expected true match and 18 false, with 27 missing matches remaining among patterns with $\mathrm{P}(M|\boldsymbol{\gamma}) < 0.5$. In comparison, the present approach has achieved better results even taking into account only the solved patterns (with 5,172 true match and only one false match). In addition, it is worth noting that filtering method probably excluded more than 900 matches from the analysis, as highlighted by the difference between the 6,037, expected by our method, and the 5,129 estimated by the sorted neighbour example. This difference would not have been detected if we had not used the proposed approach.

Table 5. Expected numbers of matched, unmatched pairs and conditional match probabilities according to patterns of agreement (solved patterns in bold).

First last names	Gender	Country id	Year of birth	Month of birth	Day of birth	M	U	$P(M\|\gamma)$	First last names	Gender	Country id	Year of birth	Month of birth	Day of birth	M	U	$P(M\|\gamma)$
1	**1**	**1**	**1**	**1**	**1**	**4157**	**0**	**1**	0	0	0	1	1	1	2	5296	0.0004
1	**0**	**1**	**1**	**1**	**1**	**220**	**0**	**0.9999**	1	0	0	0	0	1	0	349	0.0002
1	**1**	**0**	**1**	**1**	**1**	**597**	**0**	**0.9995**	0	1	1	0	1	1	5	19368	0.0002
1	0	0	1	1	1	321	0	0.9988	0	0	1	1	0	1	0	3058	0.0001
1	**1**	**1**	**1**	**0**	**1**	**61**	**0**	**0.9973**	1	0	0	0	1	0	0	932	0.0001
1	0	1	1	0	1	33	0	0.9934	0	0	0	0	0	0	0	145832310	0
1	**1**	**1**	**1**	**1**	**0**	**64**	**0**	**0.9932**	0	1	0	0	0	0	0	124788814	0
1	**1**	**1**	**0**	**1**	**1**	**73**	**1**	**0.9837**	0	0	1	0	0	0	0	7537095	0
1	0	1	1	1	0	34	1	0.9832	0	0	0	0	1	0	0	13054930	0
1	1	0	1	0	1	89	4	0.9608	0	0	0	0	0	1	0	4892021	0
1	0	1	0	1	1	39	2	0.9603	0	1	1	0	0	0	0	6449497	0
1	1	0	1	1	0	93	10	0.9057	0	0	0	1	0	0	0	1763600	0
1	1	0	0	1	1	106	27	0.7986	0	1	0	0	1	0	0	11171113	0
1	1	1	1	0	0	9	6	0.6291	0	1	0	0	0	1	0	4186106	0
1	0	0	1	0	1	5	4	0.5263	0	0	1	0	1	0	0	674722	0
1	1	1	0	0	1	11	15	0.4119	0	1	0	1	0	0	0	1509113	0
1	0	0	1	1	0	5	11	0.3031	0	0	1	0	0	1	0	252836	0
1	1	1	0	1	0	11	41	0.2152	0	0	1	1	0	0	0	91149	0
1	0	0	0	1	1	6	31	0.1522	0	0	0	0	1	1	0	437934	0
0	**1**	**1**	**1**	**1**	**1**	**26**	**205**	**0.1129**	1	0	0	0	0	0	0	10409	0
1	0	1	1	0	0	0	7	0.0713	0	1	1	0	1	0	0	577360	0

Table 5. Continued

First last names	Gender	Country id	Year of birth	Month of birth	Day of birth	M	U	$P(M\|\gamma)$	First last names	Gender	Country id	Year of birth	Month of birth	Day of birth	M	U	$P(M\|\gamma)$
0	0	1	1	1	1	14	274	0.0487	0	0	0	1	1	0	0	157878	0
1	0	1	0	0	1	1	18	0.0307	0	1	1	0	0	1	0	216352	0
1	1	0	1	0	0	1	108	0.0124	0	0	0	1	0	1	0	59161	0
1	0	1	0	1	0	1	48	0.0123	0	1	1	1	0	0	0	77996	0
0	1	0	1	1	1	38	4532	0.0083	0	1	0	0	1	1	1	374741	0
1	1	0	0	0	1	2	299	0.0052	1	1	0	0	0	0	0	8907	0
1	1	0	0	1	0	2	797	0.002	0	1	0	1	1	0	1	135096	0
0	1	1	1	0	1	4	2616	0.0015	0	0	1	0	1	1	0	22634	0
1	0	0	1	0	0	0	126	0.0006	0	1	0	1	0	1	1	50624	0
0	1	1	1	1	0	4	6982	0.0006	1	0	1	0	0	0	0	538	0
1	1	1	0	0	0	0	460	0.0004	0	0	1	1	1	0	0	8160	0

Table 6. Sorted Neighbours: marginal probability of agreement for each key variable conditional to match status of the pair.

	First and last names	Gender	Country id	Day of birth	Month of birth	Year of birth
k	1	2	3	4	5	6
m_k	0.997	0.935	0.863	0.991	0.989	0.987
u_k	0.026	0.479	0.201	0.012	0.082	0.032

The improvement of the present method if compared to the standard one becomes even more evident when further effort is carried out to consider the nine additional patterns with posterior probability $\mathrm{P}(M|\gamma) \geq 0.5$, thus, achieving a whole number of 5,445 true match and 30 false matches. However, considering that such better results are obtained by checking 510,429 pairs more than those examined for the filtered case, it is possible to limit the present method to probe only additional profiles whose $\mathrm{P}(M|\boldsymbol{\gamma}) \geq 0.9$. With this tighter limit, the pairs examined in the two cases become almost the same (851,628 against 861,628), while the gain of the present method remains significant. In fact, 5,364 linked pairs are obtained and only five of them are expected to be false, that is, 230 true matches more and 13 false matches less.

6. Final Remarks

This article introduces sampling features to probabilistic record linkage and extends the model estimation capabilities to large files while remaining in a Fellegi-Sunter framework. In this way, computational efforts were managed without affecting the set of pairs object of the study. It outperforms traditional filtering methods in that it accomplishes the estimation of linkage probabilities as to the whole set of pairs, which would otherwise be excluded from the analysis. In addition, it is effective in driving further analysis on patterns that have not been solved but that show evidence of including a high number of matches at the end of the estimation phase.

In addition to the simulations demonstrating the accuracy of the method, a real case application was presented to show how it can be used in practice for record linkage. The example is also useful to see how the method works when key variables are compared by fuzzy string similarity functions.

In our opinion, trials on larger datasets, in which filtering techniques are subject to more failures, will take advantage of this method to its fullest extent. Nevertheless, we believe that the robust approach to the estimation of linkage probabilities can be useful ‘per se’ to reduce filtering as much as possible, in such a way reducing the risk of excluding by mistake true matches from the analysis. On the other hand, our sample based filtering could be used in addition to a given standard filtering method to verify that the latter does not exclude too many “valid” pairs (matches) by mistake before the linkage step.

Further improvements include efficient sampling schemes on the set of pairs Ω to make better inferences on the contingency table of pairs by their pattern of agreement. Moreover, the use of log-linear models might improve the estimate of small frequencies for sampled patterns having small frequencies (patterns with a relatively high agreement between key variables) so as to mitigate the side effects of sampling zeros.

A better estimate of frequency distribution could also be exploited by the robust EM approach, after it is properly adjusted in order to release some of the conditional independence assumptions between key variables. In doing so, a certain degree of association between key variables could be considered in order to improve the estimate of linkage probabilities as suggested by Thibaudeau (1993).

Other developments could be related to specific strategies to adapt the probabilistic procedure to deterministic linkage steps carried out in advance. For example, a scheme that, after a deterministic linkage based on perfect agreement between a given set of key variables, applies a probabilistic step to records not linked at previous stage by using the same set of key variables, could be developed. In this case, a structural zero should be considered for patterns of perfect agreement already used during the deterministic step, and, as a consequence, an appropriate conditional probabilistic model is to be defined.

7. References

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete multivariate analysis.* Cambridge, Mass.: MIT Press. DOI: https://doi.org/10.1007/978-0-387-72806-3.

Baxter, R., P. Christen, and T. Churches. 2003. "A Comparison of Fast Blocking Methods for Record Linkage". *CMIS Technical Report 03/139*, six-pages version of the paper published in Proceedings of ACM SIGKDD '03. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.4563&rep=rep1&type=pdf (accessed April 2020).

Christen, p. 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science and Business Media. DOI: https://doi.org/10.1007/978-3-642-31164-2.

Cibella, N., M. Fortini, M. Scannapieco, L. Tosco, and T. Tuoto. 2009. "Theory and practice in developing a record linkage software". *Insights on Data Integration Methodologies*: 37–56. Available at: https://ec.europa.eu/eurostat/documents/3888793/5845197/KS-RA-09-005-EN.PDF/4cef0f2d-45a0-46b7-bfd6-196a55fca801?version=1.0 (accessed April 2020).

Cormen, T.H., C.E. Leiserson, R.L. Rivest, and C. Stein. 2009. *Introduction to algorithms*. MIT press. Available at: https://mitpress.mit.edu/books/introduction-algorithms-third-edition (accessed April 2020).

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society* B 39: 1–38. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Hernandez, M.A., and S.J. Stolfo. 1995 "The merge/purge problem for large databases". Edited by M.J. Carey and D.A. Schneider in *SIGMOD*, 127–138. DOI: https://doi.org/10.1145/568271.223807.

Herzog, T.N., F.J. Scheuren, and W.E. Winkler. 2007. *Data quality and record linkage techniques*. Springer Science and Business Media. DOI: https://doi.org/10.1007/0-387-69505-2.

Fellegi I., and A.B. Sunter. 1969. "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 328: 1183–1210. DOI: https://doi.org/10.1080/01621459.1969.10501049.

Fortini, M., L. Mancini, L.Marcone, E.Mussino, and E. Paluzzi. 2013. "Who Settles Down in Italy? Transition to Residency of non-EU Migrants". *Rivista Italiana di Economia Demografia e Statistica, no. LXVII*, (3/4). Available at: https://www.sieds.it/listing/RePEc/journl/2013LXVII_N34rieds.pdf (accessed April 2020).

Jaro, M.A. 1989. “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida”. *Journal of the American Statistical Association*, 84: 414–420. DOI: https://doi.org/10.1080/01621459.1989.10478785.

Larsen, M.D., and D.B. Rubin. 2001. “Iterative automated record linkage using mixture models”. *Journal of the American Statistical Association*, 96(453): 32–41. DOI: https://doi.org/10.1198/016214501750332956.

Murray, J. 2015. “Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering”. *Journal of Privacy and Confidentiality*, *7*(1). DOI: https://doi.org/10.29012/jpc.v7i1.643.

Neykov, N., P. Filzmoser, R. Dimova, and P. Neytchev. 2007. “Robust fitting of mixtures using the trimmed likelihood estimator”. *Computational Statistics & Data Analysis*, *52*(1): 299–308. DOI: https://doi.org/10.1016/j.csda.2006.12.024.

Newcombe, H.B., J.M. Kennedy, S.J. Axford, and A.P. James. 1959. “Automatic linkage of vital records”. *Science*, 130(3381): 954–959. DOI: https://doi.org/10.1126/science.130.3381.954.

Thibaudeau Y. 1993. “The discrimination power of dependency structures in record linkage”. *Survey Methodology*, 19: 31–38. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1993001/article/14477-eng.pdf (accessed April 2020).

Winkler, W.E. 1988. “Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage”. *Proceedings of the Section on Survey Research Methods: American Statistical Association:* 667–671. Available at: https://www.asasrms.org/Proceedings/papers/1988_124.pdf (accessed April 2020).

Winkler, W.E. 1989. ”Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage”. *Proceedings of the Fifth Census Bureau Annual Research Conference*, March 19-22, Arlington, Virginia, U.S.A.: 145–155. Available at: https://www.academia.edu/34177520/Near_Automatic_Weight_Computation_in_the_Fellegi-Sunter_Model_of_Record_Linkage (accessed April 2020).

Winkler, W.E. 2006. “Overview of record linkage and current research directions”. *Bureau of the Census Working Paper No. RRS2006-02*. Available at: https://www.census.gov/library/working-papers/2006/adrm/rrs2006-02.html (accessed April 2020).

Yancey, W.E. 2002. “Improving EM Algorithm Estimates for Record Linkage Parameters”. *Proceedings of the Section on Survey Research Methods: American Statistical Association*. Available at https://www.asasrms.org/Proceedings/y2002/Files/JSM2002-000581.pdf (accessed April 2020).

Received July 2018
Revised May 2019
Accepted July 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 827–854, http://dx.doi.org/10.2478/JOS-2020-0040

Three-Form Split Questionnaire Design for Panel Surveys

Paul M. Imbriano[1] *and Trivellore E. Raghunathan*[2]

Longitudinal or panel surveys are effective tools for measuring individual level changes in the outcome variables and their correlates. One drawback of these studies is dropout or nonresponse, potentially leading to biased results. One of the main reasons for dropout is the burden of repeatedly responding to long questionnaires. Advancements in survey administration methodology and multiple imputation software now make it possible for planned missing data designs to be implemented for improving the data quality through a reduction in survey length. Many papers have discussed implementing a planned missing data study using a split questionnaire design in the cross-sectional setting, but development of these designs in a longitudinal study has been limited. Using simulations and data from the Health and Retirement Study (HRS), we compare the performance of several methods for administering a split questionnaire design in the longitudinal setting. The results suggest that the optimal design depends on the data structure and estimand of interest. These factors must be taken into account when designing a longitudinal study with planned missing data.

1. Introduction

1.1. Motivation for Planned Missing Data Designs

Longitudinal of panel surveys are essential for any study of change in the key outcome variables and their correlates. As it is costly to conduct and recruit participants into such studies, researchers often try to get the most information they can out of study participants. For the purpose of design planning, survey costs may be decomposed into fixed and variable costs (Cochran 1977 chap. 5, 96–100; Groves 1989 chap. 2, 50–76). Fixed or overhead costs do not depend on sample size. In contrast, variable costs increase as the sample size of the study increases. Costs associated with sampling frame, sampling design, and questionnaire design are relatively fixed, while labor costs associated with interviewing and data entry, printing costs, and mailing costs are variable costs (Yansaneh 2005; Deutskens et al. 2006). Although lengthening a survey questionnaire will increase the cost of a study, the costs associated with adding an additional question to a questionnaire are generally small compared to the overall costs of the study (Deutskens et al. 2006; Groves 1989 chap. 10, 490–496).

Due to the costs of conducting a study, survey questions may be pooled from several investigators with multiple research interests, as it is generally cheaper to conduct a single large study instead of several small studies. This results in long questionnaires and

[1] Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights Ann Arbor, MI 48109-2029, U.S.A. Emails: pimbri@umich.edu, and paulmimbriano@gmail.com
[2] Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson St, Ann Arbor, MI 48104, U.S.A. Email: teraghu@umich.edu

increases the burden on participants. These studies, therefore, are also subject to dropout or nonresponse, which may lead to biased results. Sharp and Frankel (1983) found that survey length was associated with perceived burden. Furthermore, longer surveys can affect the quality of participants' responses. Several studies have shown that nonresponse rates tend to be high in surveys with long questionnaires (Adams and Darwin 1982; Dillman et al. 1993; Roszkowski and Bean 1990), and item nonresponse is more frequent towards the end of a questionnaire (Raghunathan and Grizzle 1995). Roszkowski and Bean (1990) found the response rate was 28% higher for the short version of a questionnaire compared to the long version. Past a certain length, participants become more likely to lose interest in the study, making responses less accurate (Herzog and Bachman 1981; Gonzalez and Eltinge 2007; Peytchev and Peytcheva 2017). Galesic and Bosnjak (2009) found that participants spent less time responding to items located at the end of a survey compared to items located at the beginning of the survey. Items at the end of the study also had lower response rates, shorter responses for open-ended items, and less variation in responses when items on the same response scale were arranged in a grid. A shorter survey length alleviates these problems and potentially decreases the cost of data collection per subject. The goal, therefore, is to balance collecting a set of rich variables while not placing an undue burden on participants.

The problems caused by lengthy questionnaires are exacerbated in longitudinal studies due to calling on respondents to fill out the questionnaires repeatedly. In addition to the issues of fatigue and bias from a lengthy survey, longitudinal studies must also take into consideration the effect that a lengthy survey may have on attrition. Sharp and Frankel (1983) found that, when asked if they would agree to participate in a second interview one year later, a higher proportion of participants who received the longer survey (27%) indicated that they would not agree to a follow-up interview compared to the group that received the shorter survey (13%). The actual follow-up rates one year later were closer for the long versus short survey groups (85% versus 88%), but still favored the shorter interview group. Zabel (1998) found that a planned reduction in the length of interviews for the Panel Study of Income Dynamics led to a decrease in the attrition rate.

1.2. Planned Missing Data Designs and Split Questionnaires

Meanwhile, advancements in software for handling missing data, especially the multiple imputation approach, have made missing data less problematic for data analysis, and, in fact, designing a survey to purposely include missing data could improve the quality of the study (Littvay 2009). A planned missing data design provides an effective way to reduce questionnaire length, while maintaining all relevant questions of interest. Furthermore, missing values resulting from the planned missing approach are by design either missing completely at random (MCAR) or missing at random (MAR) (Rubin 1976; Little and Rubin 2002 chap. 1, 11–13). As a result, we can use multiple imputation, maximum likelihood, or fully Bayesian approaches to handle the missing data just by focusing on the model for variables in the survey. In fact, since planned missing data designs reduce the burden on participants, the probability of nonresponse decreases, making it less likely to observe values that are missing not at random (MNAR), and, as a result, multiple imputation and maximum likelihood approaches are more likely to be valid (Rhemtulla and Little 2012; Jorgensen et al. 2014; Kaplan and Su 2016). Planned missing data approaches have frequently been used for educational assessment, where students are evaluated on several subjects. Evaluating a student's proficiency on every subject would take a great deal of time, making it disruptive for students and unlikely to be approved by administrators. For this reason, many assessments utilize either split questionnaire design or multiple matrix sampling, where each student responds to just a subset of the total questions. The Kentucky Instructional Results Information System, the Massachusetts Comprehensive Assessment System, the National Assessment of Educational Progress, and the Dutch National Assessment Program have all used multiple matrix sampling to reduce the testing burden on students (Childs and Jaciw 2002).

Raghunathan and Grizzle (1995) proposed split questionnaire design as an extension of multiple matrix sampling described in Shoemaker (1973), which randomly sampled items for each individual. Raghunathan and Grizzle (1995) modified multiple matrix sampling to place constraints on item assignment so that all population quantities of interest were estimable. Split questionnaire design divides the survey questions into multiple components and each participant responds to a fraction of the total components. One common variant, the three-form split questionnaire, divides the survey into four components (X,A,B,C). Each participant responds to all items in X and two of the three other components, resulting in three unique survey forms, (X,A,B), (X,A,C) and (X,B,C), which are administered in equal proportions (Graham et al. 2006; Rhemtulla and Little 2012). This particular three-form design reduces the survey length by approximately 25%, but modifications can be made to both the number of total components used and the fraction that each participant answers, depending on the survey composition and desired reduction in length. The split questionnaire is simple to implement and has been used in multiple studies (Graham et al. 2006).

1.3. Considerations for Implementing Split Questionnaire Designs

It is important to note that the ordering of questions within a questionnaire and where a question appears in relation to other questions could have an impact on responses (Schuman and Presser 1981; Sudman et al. 1996). Context effects should be considered during the design phase. Some context effects can be accounted for by placing variables into blocks. The questions within each block can be specifically ordered to account for context effects and the blocks themselves divided into components instead of individual variables. This should be done if certain questions are known to impact the answers to other questions. However, we might not always know if the ordering of questions will affect responses. Thus, when we design a split questionnaire, omitting certain questions or altering the ordering of questions could alter responses and possibly lead to bias. For split questionnaire designs, it would be useful to check that there are no systematic differences in responses to questions based on the assigned survey portion.

In cross-sectional studies, the split questionnaire design was found to produce estimates similar to those obtained in the absence of missing data (Raghunathan and Grizzle 1995; Littvay 2009), but with a decrease in power. The loss of statistical power can be somewhat mitigated by the increased sample size obtainable due to the decrease in cost per participant (Thomas et al. 2006; Littvay 2009). This is especially true if certain individual items are expensive to measure. Peytchev and Peytcheva (2017) also demonstrated that responses to questions from a split questionnaire design more closely resembled responses when questions are at the beginning of a lengthy survey instead of at the end, indicating that a split questionnaire may decrease bias due to fatigue. The application of a split questionnaire design should be considered when it would decrease the mean squared error (MSE) compared to a traditional survey. Unfortunately, it is difficult to know when a split questionnaire design would decrease the MSE, as the amount of bias due to burden from survey length will likely differ among studies. Although Sharp and Frankel (1983) demonstrated that an increase in survey length led to an increase in perceived burden, both the perceived usefulness of the survey and invasiveness of the questions had a larger effect on burden. A split questionnaire design could have a larger impact on the MSE for longitudinal studies if it substantially lowers the dropout rate. Fewer dropouts might decrease the variance, through an increased sample size at later study visits, and decrease nonresponse bias compared to a traditional longitudinal study.

1.4. Literature Review of Longitudinal Split Questionnaire Design

Until recently, little research had been done on the implementation of split questionnaires for longitudinal studies (Jorgensen et al. 2014). Creech et al. (2011) found that groups administered a split questionnaire had lower attrition than the groups administered the full questionnaire. Gonzalez and Eltinge (2008) and Gonzalez (2012) examined adaptively assigning items to individuals for the second interview in a panel study based on responses from the first interview. Rhemtulla and Little (2012) mentioned three strategies for assigning forms in longitudinal studies on growth curves: assigning the same forms at each visit, rotating the forms, or randomly assigning forms at each wave. It was suggested that the same forms should be given at each visit when the measure of a variable across waves is of primary interest and the forms should be rotated when examining the relationship of several latent variables. Jorgensen et al. (2014) examined longitudinal three-form split questionnaire designs in a latent variable setting. They explored the effect of longitudinal form assignment on the relative efficiency of estimating cross-lagged regression parameters (the association of a variable with another variable measured at an earlier time), autoregression parameters (the association of a variable with itself measured at different times), and factor loadings. They implemented three longitudinal form designs: assigning the same form at each visit, assigning a different form at each visit, and randomly assigning forms at each visit. In general, the same form was better at estimating cross-lagged regression and autoregression parameters, but different forms were better for factor loading. The random form assignment performed somewhere in between the same form and different form designs.

In Section 2, we propose several longitudinal designs for split questionnaires. In addition to the three assignment methods examined in Jorgensen et al. (2014), we explore several more complex methods to determine whether the more complex methods provide additional benefits in estimation compared to simpler assignment methods. In Section 3, we discuss methods to analyze data collected from a split questionnaire survey. We compare the performance of our proposed split questionnaire designs using results from simulations in Section 4 and data from the Health and Retirement Study (HRS) in Section 5. Finally, we discuss our conclusions, limitations, and areas for future research in Section 6.

2. Longitudinal Split Questionnaire Survey Designs

2.1. Issues in Longitudinal Split Questionnaire Design

Administering split questionnaires in longitudinal surveys is more complex due to repeated variable measurements on the same subject at different time points. Usually, these repeated measurements are highly correlated (Hardt et al. 2012). Also, unlike cross-sectional studies, not all variables are collected at once. A longitudinal study where participants skip certain study visits could be considered for a longitudinal split questionnaire, if the cross-sectional study questionnaire is not too burdensome. However, when the cross-sectional questionnaire is lengthy, it is more sensible for a longitudinal split questionnaire study to be comprised of cross-sectional split questionnaires administered at each visit. We might consider designing a new split questionnaire at each wave conditional on the variables already observed in the previous waves (here we define wave as any study visit in which survey questionnaires were administered to participants). However, in most studies there will be a large number of possible split questionnaire designs to choose from. Assuming we have an equal number of variable blocks in each split, there are

$$\frac{\binom{b}{q}\binom{b-q}{q}\cdots\binom{2q}{q}\binom{q}{q}}{s!} \tag{1}$$

number of potential split questionnaire designs. Here b represents the total number of variable blocks, where each block consists of one or more variables. Certain variables may be placed in the same block based on content and skip pattern. For example, a question asking, “Do you currently smoke?” may be placed in the same block as a question related to, “How many cigarettes do you smoke a day?” In the equation, s represents the number of splits or survey components into which the variable blocks are divided. Finally, q denotes the number of blocks per split. Here, we are assuming that the order of the splits does not matter. This is a reasonable assumption if there are no context effects, which have hopefully been accounted for when placing variables into blocks. With eighteen blocks and three splits, there are 2,858,856 potential split questionnaire designs to choose from. There are even more possible designs to consider if the order of splits affects responses. This can make it difficult to design a new split questionnaire at every visit. It would be simpler instead to use the same cross-sectional split questionnaire forms at each wave. Although the same forms are being used, each individual does not need to receive the same form at every visit. There are still multiple ways that we can administer the forms in a longitudinal design. We propose several methods for administering a three-form split questionnaire design in longitudinal studies.

Prior to designing a study, we need to determine whether to administer the same form to each participant throughout the entire study, rotate the form each participant receives from visit to visit, or employ some combination of those two designs, in which we administer the same form to some participants and rotate the forms for other participants. Which design works best likely depends on the correlation structure of the data, more specifically, how the correlation between components measured at the same wave (within-wave correlation) compares to the correlation of measurements on the same variable over time (autocorrelation) and the correlation between two separate components measured at different waves (between-wave correlation). When the autocorrelation is greater than the within-wave correlation, we expect it would be preferable for most estimates to not measure the same components at each wave, as the missing values would be highly correlated with the variable observed at other waves. We expect that this would improve estimation (Raghunathan and Grizzle 1995; Thomas et al. 2006). Note that in this article, we use the term autocorrelation to refer to a variable's correlation with itself over time. This should not be confused with the use of the term autocorrelation in the context of time series analysis, where autocorrelation refers to the correlation of residuals from a regression model across time points (Box and Pierce 1970). We also hypothesize that when most correlations are non-zero, more complex form rotations will produce more precise estimates by better estimating these different correlations. We will examine this hypothesis in our analysis and simulations. Which quantity is of primary interest to investigators could also influence which design should be administered. If we are primarily interested in estimating the change in a variable over time, it might be preferable to administer the same components throughout the study so that a variable can be measured in the same individuals at all time points.

2.2. *Proposed Longitudinal Split Questionnaire Designs*

We consider six different design options for allocating a three-form questionnaire in longitudinal studies. Table 1 displays how each component, A, B, and C, would be allocated under the first five proposed designs in a three-wave study, but each design could be easily modified depending on the number of waves and desired reduction in survey length. For the first two options, the participants are placed into three groups. With Option 1, each group receives either form (A,B), (A,C), or (B,C) throughout the study, while for Option 2 the forms each group receives are rotated in a manner that allows each form to be given in equal proportions at each wave. Option 1 is analogous to the same form design in Jorgensen et al. (2014). We expect it to perform better at estimating the change in a variable over time as it performed better at estimating autoregression parameters in Jorgensen et al. (2014). Option 2 is analogous to the different form design in Jorgensen et al. (2014). For Option 3, we again administer to participants a different form at each wave, but we instead create six groups, which provides more ways of cycling the forms. We hypothesize that the more complex design of Option 3 may allow us to better model variable correlations than Option 2 and will outperform it when most correlations are fairly strong. We will examine this hypothesis through our numerical analysis and simulations.

Combining aspects of the first two options may produce split questionnaire designs that perform well in estimating both the cross-sectional and longitudinal properties of variables. We would expect Option 1 to better estimate longitudinal properties, such as the change in a variable over time, as the same variables in individuals are measured at all waves. However, when autocorrelations are large, we expect that rotating the forms would perform better at estimating cross-sectional properties, such as the variable means at each wave, than administering the same form at each wave because a variable is likely more predictive of its missing value than other variables. Collecting variables that are highly

Table 1. Form allocation by design option.

Option	Wave 1	Wave 2	Wave 3
1	AB	AB	AB
	AC	AC	AC
	BC	BC	BC
2	AB	AC	BC
	AC	BC	AB
	BC	AB	AC
3	AB	AC	BC
	AB	BC	AC
	AC	BC	AB
	AC	AB	BC
	BC	AB	AC
	BC	AC	AB
4	AB	AB	AB
	AB	AC	BC
	AC	AC	AC
	AC	BC	AB
	BC	BC	BC
	BC	AB	AC
5	AB	AB	AB
	AB	AC	BC
	AB	BC	AC
	AC	AC	AC
	AC	BC	AB
	AC	AB	BC
	BC	BC	BC
	BC	AB	AC
	BC	AC	AB

predictive of missing values is helpful for multiple imputation (Collins et al. 2001; Thomas et al. 2006; Hardt et al. 2012). There may also be scenarios where one of Option 1 or Option 2 does not perform well, but a design combining aspects of the two options still performs fairly well. Option 4 represents a combination of the first and second options. For this option, there are a total of six groups and half of the participants receive the same form throughout and the other half follow the rotation for Option 2. Option 5, the most complex of the planned study designs, contains nine groups, three of which receive the same forms as in Option 1 and the other six are rotated like in Option 3.

The final design, Option 6, randomly assigns forms at each wave, making it simple to administer. This option allows every possible rotation to occur, which we believe is beneficial when most correlations are non-zero, but makes the form design unbalanced. For example, under Option 6, we no longer control the number of individuals who receive form AC after receiving AB. Random form assignment was also examined in Jorgensen et al. (2014), where it was found to perform somewhere in between the same and different form designs. Our hypothesis is that the more complex designs would perform well under many types of correlation structures and for multiple estimands, but might not necessary be optimal under any scenario or for any estimate. Depending on the performance of proposed design options and estimands of interest, this may be a worthwhile trade-off.

3. Analysis of Split Questionnaire Surveys

Due to the presence of missing data, maximum likelihood, fully Bayesian approaches, or multiple imputation are necessary for data analysis. We focus on results from data analysis using both maximum likelihood and multiple imputation for comparing the performance of our proposed split questionnaire designs.

3.1. Maximum Likelihood Estimation

Missing values resulting directly from the split questionnaire survey implementation are by design. Because of this, these missing values are either MCAR or MAR. Therefore, we can ignore the missing data mechanism and base our inference on only the observed data likelihood. For most instances, we can compute the maximum likelihood estimator (MLE) using an iterative method, such as Newton-Raphson or the EM algorithm and asymptotic standard errors can be obtained by inverting the Fisher information matrix or using other methods like bootstrapping (Little and Rubin 2002 chap. 9, 190–199).

In the case where complete data from each individual follows a p dimensional multivariate normal distribution with mean μ and variance Σ, we can obtain closed form solutions for the information matrix. Hartley and Hocking (1971) demonstrated how to compute the Fisher information of the joint likelihood of $\mathrm{Y}_1, . . . \mathrm{Y}_\mathrm{n}$ for any arbitrary missing data pattern.

For each observation, Y_i, we can construct a $p \times p$ matrix such that for $1 \leq j \leq p$ each entry $a_{j,j} = 1$ if variable j was observed and 0 otherwise. We can then take d_i equal to this constructed matrix after all rows of zeros have been deleted, creating a $o_i \times p$ matrix where o_i denotes the number of variables observed on subject i. Each observation Y_i follows a multivariate normal distribution with mean $\mathrm{d}_\mathrm{i}\mu$ and variance $\mathrm{d}_\mathrm{i}\Sigma\mathrm{d}_\mathrm{i}^\mathrm{T}$.

The expected information matrix is block diagonal in the multivariate normal case, allowing us to compute and invert the expected information for μ and Σ separately. For μ we can conveniently write the expected Fisher information for the total sample as

$$I_\mu = \sum_{i=1}^{n} \mathrm{d}_\mathrm{i}^\mathrm{T}(\mathrm{d}_\mathrm{i}\Sigma\mathrm{d}_\mathrm{i}^\mathrm{T})^{-1}\mathrm{d}_\mathrm{i} \qquad (2)$$

(Hartley and Hocking 1971). We can obtain the asymptotic variance of μ by inverting I_μ, assuming that Σ is either known or we have a consistent estimate of Σ. The variance obtained from inverting this matrix is asymptotically equivalent to the variance from using the best linear unbiased estimation (BLUE) method described by Chipperfield and Steel (2009).

Now, let σ_{jk} denote the element of Σ located at row j and column k. For the information matrix of Σ, it is convenient to create a vector of length $\binom{p+1}{2}$, δ, which contains all unique elements of Σ, or all σ_{jk} such that $j \leq k$. The information matrix based on δ is a $\binom{p+1}{2} \times \binom{p+1}{2}$ matrix. Let $\Sigma_\mathrm{i} = \mathrm{d}_\mathrm{i}\Sigma\mathrm{d}_\mathrm{i}^\mathrm{T}$ denote the variance matrix for subject i and let $\mathrm{I}_{\Sigma\mathrm{i}}$ denote the expected information from subject *i*. The formula for the expected information for subject *i*, corresponding to the negative expectation of the partial derivative, is as follows:

$$-E\left[\frac{\partial^2 \, log \, L}{\partial \sigma_{jk} \partial \sigma_{lm}}\right] = \frac{1}{2} tr\left(\Sigma_{\mathrm{i}}^{-1} \Sigma_{\mathrm{ijk}} \Sigma_{\mathrm{i}}^{-1} \Sigma_{\mathrm{ilm}}\right), \tag{3}$$

where Σ_{ijk} is a $o_i \times o_{\mathrm{i}}$ matrix with entries corresponding to elements σ_{jk} and σ_{kj} in Σ_{i} equal to one and all other elements equal zero (Hartley and Hocking 1971). Thus, $-E\left[\frac{\partial^2 \, log \, L}{\partial \sigma_{jk} \partial \sigma_{lm}}\right] = 0$ if σ_{jk} or σ_{im} are not elements of Σ_{i}. From this formula we can calculate the expected information matrix for each observation. Due to the additive property of Fisher information, the total expected information for $\mathrm{Y}_1, \ldots \mathrm{Y}_{\mathrm{n}}$ is equal to the sum of the expected information for each individual observation. Therefore,

$$\mathrm{I}_{\Sigma} = \sum_{i=1}^{n} \mathrm{I}_{\Sigma \mathrm{i}} \tag{4}$$

where I_{Σ} is the expected Fisher information for Σ. We can obtain the asymptotic variances for our MLE estimates of μ and Σ by inverting the expected information matrices, assuming that all elements of μ and Σ are estimable from our study. We can use the variances obtained from the expected information matrix to directly compare the efficiency of two proposed split questionnaire designs in estimating certain quantities. For the multivariate normal distribution, the expected information matrix does not depend on μ. We can directly compute the expected information matrix if Σ is known, which enables a comparison of two proposed split questionnaire designs when the covariance structure of the data is known. When Σ is unknown, we can iteratively solve for μ and Σ, as described by Hartley and Hocking (1971), and obtain estimates for the large-sample variance-covariance matrices of our parameter estimates based on the expected Fisher information. It is suggested that estimating the variance based on the observed information is generally preferable in the presence of missing data, as the observed information provides valid estimates for the asymptotic variance when data are MAR, while the expected Fisher information requires that the missing data mechanism is correctly specified in order to produce consistent estimates of the variance (Kenward and Molenberghs 1998; Little and Rubin 2002, chap. 11, 223–227). The expectation of the observed information should be taken with respect to the joint distribution of the data and the missing data mechanism. Ignoring the missing data mechanism when taking the expectation (as we have done here) is only valid when data are MCAR. However, for the purpose of comparing our proposed split questionnaire designs, missing data are MCAR by design.

3.2. Multiple Imputation

Multiple imputation (MI) is a frequently used method for handling missing data (Chhabra et al. 2017). With MI, we fill in missing values to create several complete data sets. The MI approach is particularly convenient because, after performing MI, standard statistical analysis can be applied on each imputed data set. The estimated parameters and variances from each analysis are combined using Rubin's rules (Rubin 1987, 76) to obtain our final estimates and variances. Often, multiple imputation is done using a Sequential Regression Multiple Imputation (SRMI) framework, where missing values are drawn using Gibbs sampling from the posterior predictive distribution of a regression model, with each missing variable regressed on all other variables (Raghunathan et al. 2001). Several popular multiple imputation software packages, like IVEWare and MICE, use the SRMI framework (Raghunathan et al. 2002; Van Buuren and Groothuis-Oudshoorn 2011).

Since imputation uses information from other variables to estimate missing values, having observed values highly correlated with missing values will better predict those missing values and improve imputations (Collins et al. 2001; Thomas et al. 2006; Hardt et al. 2012). Ideally, we would design split questionnaire surveys that take advantage of this attribute; for this reason, Raghunathan and Grizzle (1995) assign variables using correlations, where variables with high partial correlations are placed in different components. For most cross-sectional studies, we simply administer forms AB (X,A,B), AC (X,A,C), and BC (X,B,C) in equal proportions and focus our attention on how to place the variables into each component.

4. Simulation and Analysis with Proposed Split Questionnaire Designs

4.1. Comparing the Performance of Split Questionnaire Designs

We performed simulations to examine how well the proposed designs perform under a number of different correlation structures when data follow a multivariate distribution. Several papers describe methods for determining optimal split questionnaire designs. Thomas al. (2006) focused on assigning split questionnaires so observed values would be more predictive of missing values. Chipperfield and Steel (2009) considered maximizing the efficiency of estimated population totals for a fixed cost or minimizing the cost for a fixed variance. Chipperfield and Steel (2011) examined optimal split questionnaire designs with respect to costs for estimating either variable means for multivariate normal data, probabilities for multinomial data, or parameter estimates for linear regression. Chipperfield et al. (2018) examined split questionnaire designs for binary variables in terms of minimizing the loss of information for a generalized logistic regression model. Here, there is a single outcome and multiple covariates. Adiguzel and Wedel (2008) proposed minimizing the Kullback-Leibler (KL) divergence between the observed and complete data likelihoods for determining an optimal split questionnaire design. The KL divergence is a measurement of the distance between two probability distributions. The idea is to find the split questionnaire design that will produce an observed data likelihood that is closest to the complete data likelihood in the absence of missing data. All of these papers looked at optimal designs for cross-sectional studies. For the purpose of longitudinal designs, we want to include estimation of change in a variable over time. We also want to consider that there may be several regression models fit to the data and do not want to distinguish between outcome variables and covariates. Also, depending on the purpose of the study, certain estimands may be of more interest than others and this may not be captured by a single overall summary measure like KL divergence.

We examined the performance of our proposed design options in terms of how well they estimated three key quantities of interest: variable means, variance-covariance components, and the linear change in variable means over time. The estimates for the means and variance-covariance provide all the information necessary to characterize the multivariate normal distribution. For most surveys, investigators are primarily interested in estimating the population mean of a variable or variables, or in performing regression analysis. For multivariate normal data, estimated parameters from a non-repeated linear regression model can be obtained by re-parameterizing the mean and variance-covariance estimates. The design that better estimates these parameters will produce better regression estimates. One of the major reasons for conducting a longitudinal study is to assess change in a variable over time (Cook and Ware 1983). Hence, the change over time is likely of interest to investigators. We estimated our quantities of interest using both maximum likelihood estimation and multiple imputation.

4.2. Simulation Setup

For our data, we took three variables at each wave to represent the A, B, and C components of the split questionnaire with the X component omitted for simplicity, using a sample size of 108. The nine variables come from a multivariate normal distribution with the variance-covariance structure shown in Table 2. We used five parameters for the covariances, ρ_1 denoting the within-wave correlation, ρ_2 and ρ_3 the autocorrelations, and ρ_4 and ρ_5 representing the correlation between two separate components measured at different waves. The means of each variable changed linearly over time.

By varying the values of ρ_1, ρ_2, ρ_3, ρ_4, and ρ_5 we can create numerous correlation structures. Table 3. displays the correlations we used to test the performance of different options. For the first three correlation structures, only one of the within-wave, autocorrelation, or between-wave correlation is non-zero, enabling easy comparisons between the performance of each design under extremely different conditions. These first three correlation structures are not intended to represent typical correlation structures for repeated measures data, though we may find examples of variables that follow a similar correlation structure to Structure 1 and Structure 2. Say one of the variables represents a binary indicator of an infectious disease (like the flu) and the other variables represent possible symptoms related to that disease (such as fever). If enough time passes between consecutive longitudinal waves such that any acute infection will likely have resolved, we might observe a correlation similar to Structure 1. In this instance, we would expect

Table 2. Variance-covariance structure.

		Wave 1			Wave 2			Wave 3		
		A	B	C	A	B	C	A	B	C
Wave 1	A	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4	ρ_3	ρ_5	ρ_5
	B	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4	ρ_5	ρ_3	ρ_5
	C	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2	ρ_5	ρ_5	ρ_3
Wave 2	A	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4
	B	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4
	C	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2
Wave 3	A	ρ_3	ρ_5	ρ_5	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1
	B	ρ_5	ρ_3	ρ_5	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1
	C	ρ_5	ρ_5	ρ_3	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1

Table 3. Correlation structures.

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
Structure 1	0.50	0.00	0.00	0.00	0.00
Structure 2	0.00	0.50	0.50	0.00	0.00
Structure 3	0.00	0.00	0.00	0.50	0.50
Structure 4	0.80	0.50	0.50	0.40	0.40
Structure 5	0.50	0.70	0.70	0.30	0.30
Structure 6	0.50	0.70	0.49	0.25	0.125

variables collected at different waves to be weakly correlated, as disease status from the previous wave is probably not strongly related to current disease status. On the other hand, variables collected at the same wave will be correlated, as symptoms are correlated with disease status. It is easier to picture cases where variables have a similar longitudinal correlation to that of Structure 2, as this implies that data collected at the same wave are independent or weakly correlated, but a variable is moderately or strongly correlated with itself over time. We can find several variables from the data we selected from the Health and Retirement Study, a longitudinal study of US adults over 50 (Juster and Suzman 1995), similar to this structure. Additional details on the study and the selected variables can be found in Section 5. Variables like wealth, cancer diagnosis, and high blood pressure are weakly correlated with each other (correlations are between 0.024 and 0.063) but moderately to strongly correlated with themselves across waves (correlations range from 0.437 to 0.662 for wealth and from 0.947 to 0.995 for high blood pressure and cancer). Structure 3, where the between-wave correlation is the only non-zero correlation, is unlikely to be found in longitudinal data. All three structures were chosen because they allow us to test the performance of each design under radically different conditions.

The last three structures are meant to more accurately reflect correlations that might be seen in longitudinal data than the first three structures. We generally would not expect only one of the within-wave, between-wave, and autocorrelation to be non-zero. We also expect that the between-wave correlation would be no larger than the within-wave correlation and the autocorrelation. The correlations between wealth and income in the Health and Retirement Study are usually highest when collected at the same visit. For other variables, like diabetes, high blood pressure, heart disease, stroke, and weight, the between-wave correlations are usually similar to the within-wave correlation, which are both less than the autocorrelation. In Structure 4, within-wave correlation is largest, while for Structure 5 autocorrelation is largest. Structure 4 may be thought of as similar to Structure 1, but a variable is still correlated with itself and other variables when measured at different waves. We may imagine something like this occurring for certain diseases and symptoms, where the disease status is correlated over time, but is more strongly correlated with symptoms occurring during that visit. We can observe a correlation structure similar to Structure 5 with the variables income and wealth from the Health and Retirement Study, as the within-wave correlations range between 0.23 to 0.44, the autocorrelations range between 0.44 and 0.66, and the between-wave correlations range from 0.14 to 0.40. Most variables selected from the Health and Retirement Study have a higher autocorrelation than within-wave or between-wave correlation. Structure 6 follows an autoregressive structure, where correlations decrease over time, with two main correlations, the correlation between different variables and the autocorrelation. For variables like high blood pressure, diabetes, cancer, heart disease, stroke, and weight in the Health and Retirement Study, we almost always see a decline in the autocorrelation when the waves are further apart. The decline is most noticeable for heart disease and stroke.

In addition, we tested the performance of each design under a random correlation structure, with the correlation computed based on a random covariance matrix. The covariance matrix is drawn from a Wishart distribution with nine degrees of freedom and a diagonal scale matrix. The Wishart distribution, a multivariate generalization of the gamma distribution, has two parameters: the degrees of freedom, which must be greater than $p - 1$, and a $p \times p$ scale matrix, which must be positive definite. The scale matrix we used had ones on the main diagonal and zeros everywhere else. A diagonal matrix was chosen because the generated within-wave, between-wave, and autocorrelations would, on average, be the same and none of the correlations would be favored. After drawing a random covariance matrix, the covariance matrix was transformed into a correlation matrix. The generated correlation matrix has an expected value equal to the scale matrix used in the Wishart distribution, thus the generated correlations have means equal to zero. The degrees of freedom affect the variance of the generated correlation matrix. Specifying a large number of degrees of freedom would have resulted in very little variation in the generated correlations (i.e., all correlations would have been close to zero). We specified a small number for the degrees of freedom, generating a large range of correlations. The random correlations produced typically ranged from -0.65 and 0.65. The use of randomly generated correlation matrices allowed us to examine which design performs the best on average across a large number of possible correlations.

We then compared performance of each design option under the proposed correlation structures by computing the variance for the variable means, variance-covariance components, and the linear change in means over time using both MLE and mutliple imputation. For maximum likelihood estimation, we calculated the variance of the means and variance-covariance components from inverting the Fisher Information using the true underlying covariance matrix. We estimated the linear change in means over time for a variable using contrasts. The variance of these linear combination of means from the contrasts can similarly be computed directly from the inverted Fisher Information matrix. For MLE, the variances of our estimates of interest did not require data simulation to compute, though we used many iterations to assess the overall performance of Option 6, the random form assignment.

For multiple imputation, we simulated complete data from the multivariate normal distribution. Values were then set to missing so observed data matched what would have been obtained under each study design option and we performed multiple imputation. We then estimated the mean, variance- covariance, and linear change in mean over time. The mean and variance-covariance parameters were estimated from the sample mean and sample covariance matrix in each imputed data set and combined using Rubin's rules. We used a linear mixed model of the variable regressed on time for estimating the change in mean and stored the parameter estimates and standard errors for the slope and intercept.

4.3. Simulation Results

The relative performance, rankings, and conclusions are largely the same when using either MLE or MI, though there is an increase in variance from the MI estimates compared to the MLE. This might be due to the finite number of imputations being used and because we assumed a known variance-covariance matrix when computing the MLE. The MI results are based on the estimated variance-covariance parameters from imputed data, which adds an additional source of variation. Since MI is more commonly used, this section will focus on simulation results using multiple imputation. The results from MLE can be found in the supplementary materials, section S1. The results in this article are based on a longitudinal data structure with three waves. For comparison, we also performed MLE for the same basic correlation structures but with two waves. The results for two waves (not shown in this article) are similar to the results for three waves.

Table 4 displays the average percent increase in variance for the mean and variance-covariance components from using the proposed split questionnaire design versus complete data. Since relative rankings for the mean and variance-covariance were very similar, we combined them into one category in the table. Similarly, Table 5 shows the variance increase for repeated measure regression for the correlation structures. All of the design options performed similarly for Structure 1, where only variables within the same wave are correlated. The longitudinal selection of split question forms did not matter very much due to the lack of correlation across waves. However, Option 5 and Option 6 did perform slightly better at estimating variable correlations due to the additional form rotations, even if most correlations were zero.

Results from Structure 2, where only the autocorrelations are non-zero, show that Option 3 performed best in terms of estimating variable means and variance-covariance

Table 4. Average percent increase in variance from complete data for mean and variance-covariance components using MI.

Structure	1	2	3	4	5	6	Random
Option 1	90.7	134.0	109.7	38.0	67.6	79.7	68.3
Option 2	90.4	92.2	111.2	32.3	46.3	37.7	66.7
Option 3	86.2	87.9	52.8	27.7	30.1	34.3	53.3
Option 4	86.9	99.9	35.5	29.7	40.3	45.0	53.7
Option 5	85.9	94.1	33.4	28.0	32.4	39.3	49.0
Option 6	86.6	92.0	18.6	28.7	30.8	37.0	50.0

Table 5. Average percent increase in variance from complete data for repeated measures regression change in mean over time for MI.

Structure	1	2	3	4	5	6	Random
Option 1	51.5	73.0	89.7	24.1	42.0	44.6	44.3
Option 2	52.9	102.1	50.5	41.2	73.5	61.7	44.7
Option 3	52.8	103.0	31.2	36.3	50.1	58.9	37.4
Option 4	51.9	82.8	28.4	30.2	39.3	45.5	36.5
Option 5	51.2	90.7	23.0	30.5	36.6	50.2	33.4
Option 6	53.9	94.4	12.2	33.7	46.6	57.4	34.4

components. Option 2 and Option 6 were virtually tied for second, followed closely by Option 5. Option 1 performed the worst in estimating means and variance-covariance. The opposite occurred for estimating the change in mean over time. In this scenario, a variable is predictive of its values measured at different time points for the same subject. Since Option 2 and Option 3 rotate split questionnaire forms for all individuals, we measure each variable on a subject during the study, which enables us to better estimate missing values for that variable and leads to a better estimate for the mean and variance. On the other hand, Option 1 measures the same variables on an individual for every wave, which provides a larger sample of individuals with a variable measured at all time points than the other options, allowing a better estimate for how a variable changes over time. Option 4 and Option 5, which rotate split questionnaire forms for some individuals but not others, do not perform as well as Option 3 for estimating means and variance-covariance or as well as Option 1 for estimating the change in mean. However, the efficiency loss is not bad. They might be more useful for jointly estimating the means, variance-covariance, and change in means.

For correlation Structure 3, where a variable is only correlated with other variables measured at different waves, Option 6, the random form assignment, performed the best overall by far. Option 5 was the second best under this correlation, while Option 1 and Option 2 performed terribly. This indicates that extra form rotations and more complex designs are beneficial when between-wave correlations are large. For the random variable correlation, Option 5 and Option 6 again performed the best and Option and Option 2 did the worst. Based on this, it appears that the more complex designs perform better across all possible variable correlations; however, all correlation structures are not equally likely for longitudinal studies.

For the more realistic correlation structures (Structures 4, 5, and 6), we still see a fair amount of variability in the performance of the design options. The designs perform more similarly for correlation Structure 4, where the within-wave correlation is larger than the other correlations. Option 3 performed the best in terms of estimating the mean, variance, and covariance across correlation Structures 4, 5, and 6. Options 5 and 6 were fairly close behind. For these structures, within-wave and autocorrelation were greater than between-wave correlation, which is likely why Option 3 outperformed Option 5 and Option 6 in estimating the sufficient statistics. Option 3 likely outperformed Option 2 because the extra rotations allowed better estimation of the between-wave correlations. Option 1 was always the worst at estimating the mean and variance-covariance, but, generally, was the best at estimating the change in mean over time; one exception being correlation Structure 5, where Option 4 and Option 5 outperform Option 1, though not by much. This indicates that there are certain instances where we might not prefer Option 1 for measuring the change in a variable over time. Option 4 and Option 5 were usually variables at each wave on a subset of study participants. Those two options are also quite a bit better at estimating the sufficient statistics compared to Option 1 and, as a result, may be preferable.

From the analysis we see that the optimal design for a longitudinal study with planned missingness depends on the structure of the data and the estimates of interest. If the variables have high autocorrelations, we prefer Option 2 or Option 3, where participants change forms every year, for estimating the mean and variance-covariance parameters, but Option 1 would be better for estimating the change in mean over time. If the data follow a more unusual structure, it would be better to use Option 5 or Option 6 which include more form rotations. For more realistic correlations, Option 3 appears to perform best for estimating mean and variance-covariance and Option 1 is likely best at estimating change in mean. Option 4 and Option 5 can measure both mean and variance-covariance and change in mean fairly well. While our simulations are useful for evaluating how each option performs, our data structures are more simplistic than what we would observe in practice. Next, we examine which planned missing design option works best using data from a longitudinal survey.

5. Health and Retirement Study (HRS)

5.1. Overview of Selected HRS Data and Assigned Splits

The Health and Retirement Study (HRS), beginning in 1992, is an ongoing longitudinal study of US adults age 50 and older (Juster and Suzman 1995). HRS collects information pertaining to the health, income, and job status of participants. New individuals enter the study after turning 50, while some older participants exit due to death or loss to follow up. HRS uses multistage sampling to sample households (Health and Retirement Study 2008). African Americans, Hispanics, and residents of the state of Florida are oversampled compared to the general population. Sampling weights are used to account for the unequal probability of selection. The weights include adjustments for post-stratification factors. There are numerous publications analyzing data collected from HRS. An examination of recent publications indicates analyses were frequently conducted using regression models, including linear regression (Carr et al. 2018), logistic regression (Lee et al. 2017; Pavela et al. 2018; Shah et al. 2018), Cox proportional hazards model (Shah et al. 2018), and Poisson regression (Wagner and Olson 2018). Most of the regression analyses used the sampling weights to adjust for the unequal probability of selection due to the complex sampling design. Parameter estimates and standard errors were generally reported from a weighted regression model. All of these analyses can still be performed under a split questionnaire design using multiple imputation, though, in this article we did not account for the complex sampling design.

For evaluating the performance of our proposed split questionnaire design options, we selected seven modules from the survey data collected in the 2004, 2006, 2008, and 2010 waves of the study. We primarily focused on health data from the Health and Retirement Study. The modules for diabetes and blood pressure represented Component A, the heart disease and stroke module plus the weight module represented Component B, and the cancer module and the income and wealth module represented Component C of our split questionnaire design. In addition, we selected basic demographic information (age, gender, race, height, education) and past health behavior information (smoking and drinking history/status) to represent the X component measured in all participants. Several previous publications had studied the relationship between at least two of the five selected modules (diabetes, blood pressure, heart disease and stroke, weight, and income and wealth) using HRS data (Bowen 2010; Best et al. 2005; Avendano and Glymour 2008). Cancer was also selected as it is an important health condition. A total of six longitudinal modules were selected so we could divide the modules evenly into three splits.

Each module contained multiple variables related to the main condition of interest that we selected, but, for the purpose of our analyses, we focused on only a few main variables in these modules (i.e., have you ever been diagnosed with diabetes, blood pressure, heart disease, stroke, or cancer, what is your current weight, the net value of all assets, or amount of income). The survey length (as measured by the number of questions) at each follow-up wave would have been reduced by 36.7%, 23.3%, and 40.0% compared to the complete questionnaire for participants administered Component AB, AC, and BC, respectively. Although AC results in the smallest reduction of survey length, AC does not include weight, potentially making it less invasive and costly compared to the other components.

The within-wave correlations, between-wave correlations, and autocorrelations differ by variable. However, in general the autocorrelations are the largest. For diabetes, blood pressure, cancer, and weight, autocorrelations were greater than 0.85. For heart disease, autocorrelations were between 0.70 and 0.90. For stroke, autocorrelations were between 0.50 and 0.85. Autocorrelations were weaker for income and wealth, but were still between 0.30 and 0.70. Most within-wave correlations were between -0.30 and 0.30, with the exception of the correlations between wealth and income, which could get as high as 0.43. The between-wave correlations were usually very similar to within-wave correlations.

We consider all participants with no missing values in each of the seven modules for all four study waves as our complete data (ideal), a total of 3,059 subjects. The survey questions from the X module were taken from baseline only. We then set values to missing, mimicking the data pattern that would have been observed had we used the proposed planned missing data designs from Table 1. We also included Option 6, the random form assignment.

5.2. *Analysis and Results from HRS Data*

We performed multiple imputation on the missing values using MICE (Van Buuren and Groothuis- Oudshoorn 2011). The same imputation model was used for each option. Linear regression models were used to impute normally distributed variables and logistic regression models were used to impute binary variables. A variable was imputed conditional on itself at previous waves along with other health conditions and demographic data. A variable was not necessarily imputed conditional on all selected variables due to issues with collinearity for all regression models and separation of data points in logistic regression. However, care was taken to make sure the imputation model was consistent with the model later used to analyze the variables. If a variable was transformed in the later analysis, then it was transformed in the imputation. Only the main variables from each module that were used in our analyses were included in the imputation models.

To assess the validity of the imputations, we examined the marginal distribution of the imputed values compared to the observed values and checked for any major discrepancies between the two. Since data are MCAR by design, we would not expect the marginal distribution of imputed values to differ from the observed values. Furthermore, we performed diagnostics on the regression models through goodness of fit testing and diagnostic plots to examine the validity of our imputation models. For linear regression models, we plotted the residuals versus the fitted values, shown in Figure 1. If the imputation model was correctly specified then we would expect to see a random scatter

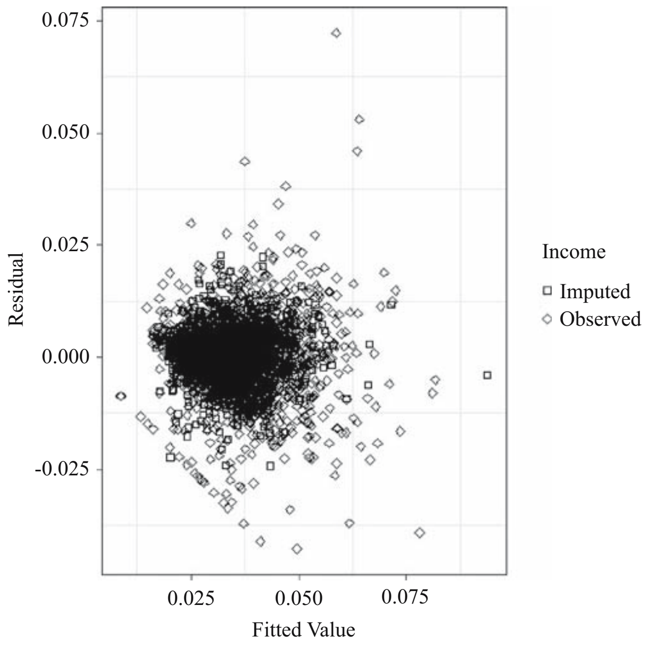


Fig. 1. Imputation diagnostic plot of residuals versus fitted values for transformation of 2010 income.

centered around the X-axis, with no discernible difference between observed and imputed values, like in Figure 1.

After performing multiple imputation, we analyzed estimates for the marginal distributions of the variables by examining the estimated mean, median, and quartiles of the continuous variables under the different options. For categorical variables we estimated π_j, the probability that variable j is equal to one. For each population parameter estimate, we examined the bias under each proposed design by taking the difference between the estimates under the planned missing designs and the estimates from the complete data. The differences for the mean and quantile estimates were standardized by dividing by the complete data standard deviation of the variables, while the difference in π_j was standardized by dividing by the square root of π_j (1 — π_j) under complete data. Figure 2 plots the distribution of the standardized differences under the six options. Option 3 produced the best estimates for univariate parameters with the bias distributed tightly around zero, with Option 2 coming in a close second. The other options performed similarly, though Option 1 performed slightly worse. In addition to examining the bias of estimated parameters, we also examined the performance of each design based on how the multiple imputation variance of the estimated population mean under each split questionnaire design compared to the variance with complete data. Figure 3 displays the ratio of the standard error of the mean estimate for each option to the standard error from the complete data. The ratio of the standard error demonstrates how much larger the standard error for the split questionnaire design is compared to the complete data. This is similar to the idea of estimating the fraction of missing information for multiple imputation (Little and Rubin 2002 chap. 10, 211), which gives an estimate for how much

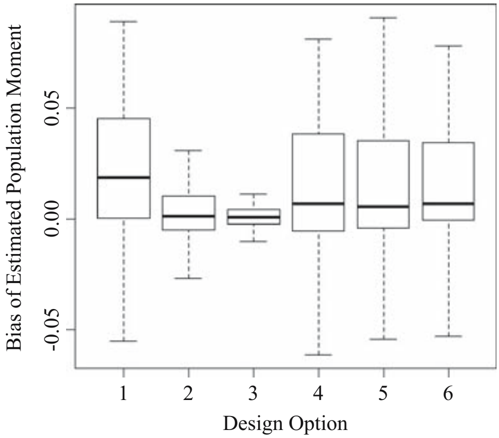


Fig. 2. Distribution of standardized bias for univariate estimates.

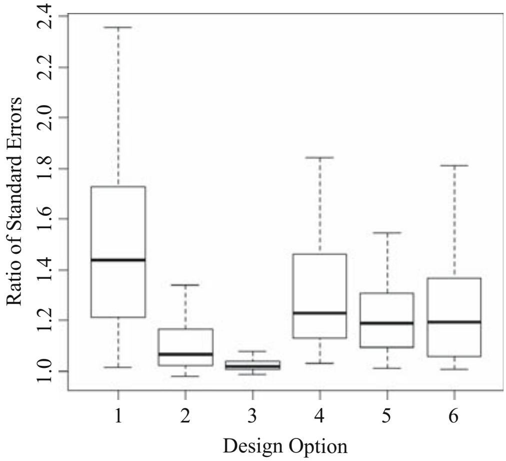


Fig. 3. Ratio of the standard error of the estimated mean to complete data standard error.

the variance of an estimate has increased due to missing data. In this instance, we do not need to estimate the fraction of missing information because we can directly determine the loss of efficiency. Once again, Option 3 performed best, generally producing smaller standard errors, followed by Option 2, while Option 1 performed the worst.

We also evaluated the performance of each option using three different regression analyses based on three previous publications with HRS data, (Bowen 2010; Best et al. 2005; Avendano and Glymour 2008). More details on the regression models and results based on these papers can be found in the supplementary materials, Section S2.

Figure 4 shows how each option performed in terms of parameter estimation bias across the three regression models, by taking the difference in parameter estimates under the design option and the complete data and dividing by the complete data standard error. Option 3 performed the best at estimating regression coefficients. Option 2 was second

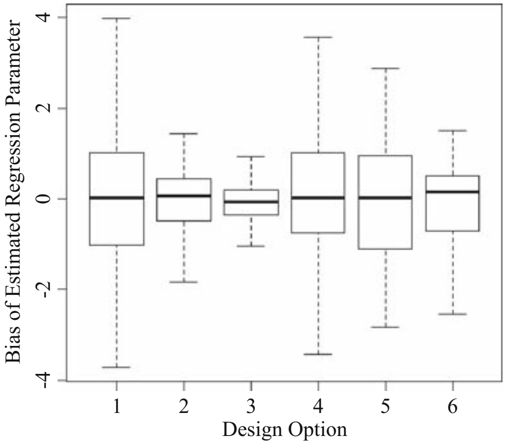


Fig. 4. Distribution of standardized regression parameter estimation bias.

best, followed by Option 6. The other three options performed similarly overall in terms of bias, though Option 1 was generally worse. In Figure 5, we display the ratio of the MI standard errors under each option to the complete data standard errors. From the figure, we observe that Option 3 produced the smallest standard errors. The other options performed fairly similar to each other, though the average standard error is slightly smaller for Option 2 and Option 6 compared to the other options.

Option 3 easily performed the best overall for the HRS data. Option 2 was the second best, while Option 1 performed the worst. The strong performance of Option 2 and 3 in this instance is not surprising since the autocorrelation was quite large for most variables, and we saw from simulations that Option 2 and Option 3 performed better in terms of mean and variance estimation when the autocorrelation is much greater than the other correlations. Had the within-wave and between-wave correlations been stronger, Option 5 and Option 6 would likely have performed better.

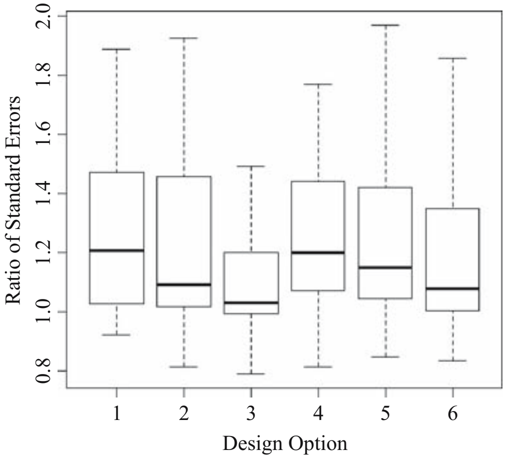


Fig. 5. Distribution of the ratio of regression parameter standard error to complete data standard error.

6. Discussion

6.1. Conclusions

The use of planned missing data in a survey study can reduce the burden and fatigue on participants, leading to an increase in the quality of the data and a reduction in unplanned missing values. This could also reduce the dropout rate in longitudinal studies. There is a good deal of literature on implementing planned missing designs in cross-sectional studies, commonly using a three-form design, but until recently there has been little research on implementing planned missing data designs in longitudinal studies, though Jorgensen et al. (2014) did consider multiple assignment methods for three-form survey designs of longitudinal studies in a latent variable setting.

In our article, we examined six different design options for allocating forms in a longitudinal study. We observed from our simulations and MLE analysis that the performance of each design option depended on the correlation structure of the data and which estimands were of interest. We also tested these designs on survey data from the HRS. To truly determine which design is the best, we need to understand the longitudinal correlation structure of the data. We could wait until we have at least two waves of data before applying a planned missing data design to the longitudinal study. However, the optimal design mostly depends on the between-wave correlation, within-wave correlation, and autocorrelation. We can estimate the within-wave correlations from the first wave before applying a longitudinal design. The between-wave correlations are unlikely to be larger than the within-wave correlations, based on what we observed in the HRS data. Thus, we only need to judge whether the autocorrelation is larger than the within-wave correlation and which estimands are of interest to determine which design to apply.

If investigators are primarily concerned with estimating the mean, regression of one variable on a subset of variables, or cross-sectional properties of the data, we recommend implementing Option 3. Option 3 always outperformed the best for most longitudinal studies due to large autocorrelations. Option always outperformed the simpler Option 2, even though both designs similarly administered a different form to a participant at every visit. Sometimes, Option 3 did considerably better than Option 2. The only reason to use Option 2 over Option 3 is that it requires fewer group assignments, making it easier to implement. Option 3 performed the best, or close to it, in terms of estimating the means and variance- covariance for every correlation structure except for Structure 3, where between-wave correlations were the only non-zero correlation, and the random correlation structure. However, the autocorrelation and within-wave correlations are likely to be stronger than between-wave correlations for longitudinal data. In these situations Option 3 was the best.

The more complex design options (Options 3, 4 and 5) generally outperformed the simpler options (Options 1 and 2) in terms of estimating the means and variance-covariance in our simulations, except for correlation Structures 2 and 6, where Option 2 outperformed Option 4 and Option 5. In these instances, the autocorrelation was quite strong compared to the between-wave correlation. The autocorrelations were quite large for the HRS data as well, which may be the reason why Option 2 performed better; however, Option 3 was still better than Option 2.

If investigators are only interested in the change in a variable over time, which may often be the case for longitudinal studies, Option 1 will usually be the best design option. This is a direct result of measuring the same variables on a participant at each time point. This result is consistent with results from Jorgensen et al. (2014), which found that the same form produces more precise estimates for autoregressive parameters. Correlation Structures 3 and 5 and the random correlation structure were the only instances where Option 1 did not perform the best or very close to the best in terms of estimating the change over time, though it was not that far behind the best option for Structure 5. Structure 5 had moderately strong with-wave and between-wave correlations and strong autocorrelations. The fairly strong overall correlations may have helped the more complex options outperform Option 1 in this instance.

Even if the change in variables over time is of primary interest, we still recommend considering Option 4 and Option 5. Option 1 was generally the worst design in terms of estimating the mean, variance, and covariance. Additionally, Option 1 did not perform well for the HRS data. Option 4 generally performed similarly to Option 1 in terms of estimating change in mean over time, but performed better in estimating the mean and variance-covariance. This may be a good design to use if interest lies in estimating the mean and variance-covariance, in addition to the change in mean. The ratio of participants receiving the same form and receiving different forms could also be altered for Option 4 and Option 5, depending on which estimand is of greater interest. Seeing how Option 3 always outperformed Option 2, redesigning Option 5 so that half of the participants received the same form would likely perform similar to Option 4 in terms of estimating the change over time, but would be better for estimating the mean and variance-covariance. Depending on the strength of the overall correlations, Option 4 and Option 5 may be better than Option 1 in terms of estimating the change in mean over time. Even though the random form assignment performed fairly well in most instances, we would recommend using Option 5 instead, as it allows us to measure the same variables every year in a subset of participants. Measuring the same set of variables in at least a subset of participants is useful for estimating the change in a variable over time.

It is also important to keep in mind that when selecting design options, certain higher-order interactions may not be directly estimable, which is also an issue for three-form split questionnaire designs in cross-sectional studies. For example, research suggests that how much influence a person believes they have over events in their lives and a person's control over his or her work environment may modify the effect of stress on health (Meier et al. 2008). If it is of interest to investigate a higher order interaction, such as the example above, care must be taken when designing the split questionnaire. Although all two-way interactions can be estimated using the three-form design, not all three-way or higher-order interactions can be estimated. Any interaction that includes a variable in component A, a variable in component B, and a variable in component C is not directly estimable. We also have to consider that some interactions involving components measured at different time points might not be estimable with the longitudinal designs. For example, Option 2 and Option 3 do not allow a three-way interaction of a variable measured in the same component across all three waves, since that variable is never measured in the same individuals in all three waves. Options 1, 4 and 5 do not have this problem. It should be taken into account during the design stage if certain higher-order interactions are of importance.

6.2. Limitations and Future Research

We evaluated the proposed design options using simulations and HRS data. Although the HRS data contained both binary and continuous variables, we did not examine any joint distributions of variables other than multivariate normal for our simulations. It would be useful to examine whether the conclusions from our simulations are affected by different variable distributions. We might want to consider instances where we have binary, categorical, or count variables in addition to the continuous multivariate normal distribution. Although, in principle, we would not expect a huge difference from the multivariate normal case (as we saw with the results from the HRS data), the distribution of our variables affects the imputation models and could result in different conclusions.

Prior information regarding the data structure would help in determining the optimal design option for a study and can help in the imputation. Several papers have considered dividing the survey questions into forms based on correlation structures or other methods to improve estimation, but require prior information on the variables (Raghunathan and Grizzle 1995; Thomas et al. 2006; Adigiizel and Wedel 2008). These methods were not considered when constructing our cross-sectional split questionnaire forms. Using these methods could potentially affect which of our longitudinal design options performs best. The results of the longitudinal split questionnaire designs were affected by cross-sectional item assignment in Jorgensen et al. (2014). We should also be aware that our results when using planned missing designs will be biased if we do not specify a correct imputation model, which by no means is a trivial matter. Even when data are MCAR, there are potential issues with collinearity, separation of data points, and violations of parametric distributions for conditional regression imputation models.

Our proposed longitudinal split questionnaire designs use the same basic three-form split questionnaire throughout the study. We could create new split questionnaire forms at every visit based on the observed data from previous visits. We could consider adaptive allocations of forms, where a participant's form at the next visit is based on previous responses. The large number of potential split questionnaires to consider at every visit makes this approach more complicated. Our proposed designs are much easier to implement. It would be a good idea to examine how the efficiency of our proposed designs compares to this approach.

One further limitation is that we did not consider the effect that complex survey designs might have on our proposed methods. Simple random sampling was implicitly assumed for our analyses; however, most large-scale studies use complex survey designs, such as multistage sampling (Zhou et al. 2016). For stratified and/or clustered sampling, our results and conclusions for administering split questionnaire designs should hold within each strata and/or primary sampling units (PSUs). For the multivariate normal distribution, overall estimates for the mean and regression parameters could be obtained by taking a weighted average of the parameter estimates within each group defined by the strata and/or PSUs, assuming we have enough observations within each group (Dumouchel and Duncan 1983). It would likely be preferable to assign split questionnaires within groups and, in some circumstances, it might be useful to apply different longitudinal designs in different groups.

When there is a small number of observations per group or our study involves post-stratification, we might use survey weights instead of computing parameters within groups. In the multivariate normal case, weighted least squares could be used for estimating regression parameters (Dumouchel and Duncan 1983). Means could be computed as a weighted average. Both the means and regression parameters can be obtained by applying transformations to the data using the weights. Our conclusions for longitudinal split questionnaire designs would likely still hold, provided the transformations do not alter the relative strengths of the within-wave and between-wave correlations and the autocorrelations from the unweighted analyses. A more thorough investigation of the effect of complex survey designs on longitudinal split questionnaire designs should be performed.

We note that though our proposed methods for longitudinal split questionnaire designs differ from rotating panel surveys and the wave missing design described in Little and Rhemtulla (2013), they are motivated by a similar concern of reducing respondent burden in longitudinal studies. In rotating panel surveys, participants are only enrolled in a panel survey for a finite period of time and new participants are enrolled in the study at each wave to replace the exiting participants. In wave missing designs, not all participants are interviewed at every wave of the study. Both designs reduce the interview burden on participants. Each design could be used in conjunction with our split questionnaire design methods to further reduce burden for longitudinal studies.

7. References

Adams, L.L.M., and G. Darwin. 1982. “Solving the Quandary Between Questionnaire Length and Response Rate in Educational Research.” *Research in Higher Education* 17(3): 231–240. DOI: http://dx.doi.org/10.1007/BF00976700.

Adigüzel, F., and M. Wedel. 2008. “Split Questionnaire Design for Massive Surveys.” *Journal of Marketing Research* 45(5): 608–617. DOI: https://dx.doi.Org/10.1509/jmkr.45.5.608.

Avendano, M., and M.M. Glymour. 2008. “Stroke Disparities in Older Americans: Is Wealth a More Powerful Indicator of Risk than Income and Education?” *Stroke* 39(5): 1533–1540. DOI: https://dx.doi.org/10.1161/STRCIKEAHA.107.490383.

Best, L.E., M.D. Hayward, and M.M. Hidajat. 2005. “Life Course Pathways to Adult-Onset Diabetes.” *Social Biology* 52(3–4): 94–111. DOI: https://dx.doi.org/10.1080/19485565.2005.9989104.

Bowen, M.E. 2010. “Coronary Heart Disease from a Life-Course Approach: Findings from the Health and Retirement Study, 1998–2004.” *Journal of Aging and Health* 22(2): 219–241. DOI: https://dx.doi.org/10.1177/0898264309355981.

Box, G.E.P., and D.A. Pierce. 1970. “Distribution of Residual Autocorrelations in Autoregressive- Integrated Moving Average Time Series Models.” *Journal of the American Statistical Association* 65(332): 1509–1526. DOI: https://dx.doi.org/10.2307/2284333.

Carr, D.C., S. Ureña, and M.G. Taylor. 2018. “Adjustment to Widowhood and Loneliness Among Older Men: The Influence of Military Service.” *Gerontologist* 58(6): 1085–1095. DOI: doi.org/10.1093/geront/gnx110.

Chhabra, G., V. Vashish, and J. Ranjan. 2017. “A Comparison of Multiple Imputation Methods for Data with Missing Values.” *Indian Journal of Science and Technology* 10(19). DOI: https:/dx.doi.org/10.17485/ijst/2017/v10i19/110646.

Childs, R.A. and A.P. Jaciw. 2002. "Matrix Sampling of Items in Large-Scale Assessments." *Practical Assessment, Research and Evaluation* 8(16). DOI: https://dx.doi.org/10.7275/gwvh-4z51.

Chipperfield, J.O., M.L. Barr, and D.G. Steel. 2018. "Split Questionnaire Designs: Collecting Only the Data that You Need through MCAR and MAR Designs." *Journal of Applied Statistics* 45(8): 1465–1475. DOI: https://dx.doi.org/10.1080/02664763.2017.1375085.

Chipperfield, J.O., and D.G. Steel. 2009. "Design and Estimation for Split Questionnaire Surveys." *Journal of Official Statistics* 25(2): 227–244. DOI: https://dx.doi.org/10.1.1.894.1568&rep=rep1&type=pdf.

Chipperfield, J.O., and D.G. Steel. 2011. "Efficiency of Split Questionnaire Surveys." *Journal of Statistical Planning and Inference* 141(5): 1925–1932. DOI: https://dx.doi.org/10.1016Zj.jspi.2010. 12.003.

Cochran, W.G. 1977. *Sampling Techniques*. New York: John Wiley & Sons, Inc. 3rd ed.

Collins, L.M., J.L. Schafer, and C.M. Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6(4): 330–351. DOI: http://dx.doi.org/10.1037/1082-989X.6.4.330.

Cook, N.R., and J.H. Ware. 1983. "Design and Analysis Methods for Longitudinal Research." *Annual Review of Public Health* 4: 1–23. DOI: https://dx.doi.org/10.1146/annurev.pu.04.050183.000245.

Creech, B., M. Smith, J. Davis, L. Tan, N. To, S. Fricker, and J.M. Gonzalez. 2011. *Measurement Issues Study Final Report. BLS Internal Report*. Available at: https://www.bls.gov/cex/research_papers/pdf/cesrvmeth_davis.pdf (accessed March 2019).

Deutskens, E., A. Jong, K. de Ruyter, and M. Wetzels. 2006. "Comparing the Generalizability of Online and Mail Surveys in Cross-National Service Quality Research." *Marketing Letters* 17: 119–136. DOI: https://dx.doi.org/10.1007/s11002-006-4950-8.

Dillman, D., M.D. Sinclair, and J.R. Clark. 1993. "Effects of Questionnaire Length, Respondent- Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys." *Public Opinion Quarterly* 57(3): 289–304. DOI: https://dx.doi.org/10.1086/269376.

Dumouchel, W.H., and G.J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78(383): 535–543. DOI: https://dx.doi.org/10.1080/01621459.1983.10478006.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349–360. DOI: http://10.1093/poq/nfp031.

Gonzalez, J.M. 2012. The Use of Responsive Split Questionnaires in a Panel Survey. PhD diss. University of Maryland. Available at: https://drum.lib.umd.edu/handle/1903/13171 (accessed March 2019).

Gonzalez, J.M., and J.L. Eltinge. 2007. "Multiple Matrix Sampling: A Review." In Proceedings of the Section on Survey Research Methods: American Statistical Association, July 29, 2007, 3069–3075. Alexandria, VA: American Statistical

Association. Available at: http://www.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000494.pdf (accessed October 2015).

Gonzalez, J.M., and J.L. Eltinge. 2008. "Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey." In Proceedings of the Section on Survey Research Methods, American Statistical Association, August 3–7, 2008, 2081–2088. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2008/Files/301351.pdf (accessed March 2017).

Graham, J.W., B.J. Taylor, A.E. Olchowski, and P.E. Cumsille. 2006. "Planned Missing Data Designs in Psychological Research." *Psychological Methods* 11(4): 323–343. DOI: https://dx.doi.org/10.1037/1082-989X.11.4.323.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Hardt, J., M. Herke, and R. Leonhart. 2012. "Auxiliary Variables in Multiple Imputation in Regression with Missing X: A Warning Against Including too many in Small Sample Research." *BMC Medical Research Methodology* 12(1): 184. DOI: https://dx.doi.org/10.1186/1471-2288-12-184.

Hartley, H., and R. Hocking. 1971. "The Analysis of Incomplete Data." *Biometrics* 27(4): 783–823. DOI: https://dx.doi.org/10.2307/2528820.

Health and Retirement Study. 2008. *Sample Evolution: 1992–1998*. Ann Arbor, MI: Institute for Social Research, University of Michigan. Available at: http://hrsonline.isr.umich.edu/sitedocs/surveydesign.pdf (accessed March 2019).

Herzog, A.R., and J.G. Bachman. 1981. "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly* 45(4): 549–559. DOI: https://dx.doi.org/10.1086/268687.

Jorgensen, T.D., M. Rhemtulla, A. Schoemann, B. McPherson, W. Wu, and T.D. Little. 2014. "Optimal Assignment Methods in Three-Form Planned Missing Data Designs for Longitudinal Panel Studies." *International Journal of Behavioral Development* 38(5): 397–410. DOI: https://dx.doi.org/10.1177/0165025414531094.

Juster, F.T., and R. Suzman. 1995. "An Overview of the Health and Retirement Study." *Journal of Human Resources* 30: S7–S56. DOI: https://dx.doi.org/10.2307/146277.

Kaplan, D., and D. Su. 2016. "On Matrix Sampling and Imputation of Context Questionnaires with Implications for the Generation of Plausible Values in Large-Scale Assessments." *Journal of Educational and Behavioral Statistics* 41(1): 57–80. DOI: https://dx.doi.org/10.3102/1076998615622221.

Kenward, M.G., and G. Molenberghs. 1998. "Likelihood Based Frequentist Inference When Data Are Missing at Random." *Statistical Science* 13(3): 236–247. DOI: https://dx.doi.org/10.1214/ss/1028905886.

Lee, M., M.M. Khan, and B. Wright. 2017. "Is Childhood Socioeconomic Status Related to Coronary Heart Disease? Evidence from the Health and Retirement Study (1992–2012)." *Gerontology & Geriatric Medicine* 3: 1–9. DOI: https://dx.doi.org/10.1177/2333721417696673.

Little, T.D., and M. Rhemtulla. 2013. "Planned Missing Data Designs for Developmental Researchers." *Child Development Perspectives* 7(4): 199–204. DOI: https://dx.doi.org/10.1111/cdep. 12043.

Little, R.J.A., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. 2nd ed.

Littvay, L. 2009. "Questionnaire Design Considerations with Planned Missing Data." *Review of Psychology* 16(2): 103–113.

Meier, L.L., N.K. Semmer, A. Elfering., and N. Jacobshagen. 2008. "The Double Meaning of Control: Three-Way Interactions Between Internal Resources, Job Control, and Stressors at Work." *Journal of Occupational Health Psychology* 13(3): 244–258. DOI: https://dx.doi.org/10.1037/1076-8998. 13.3.244.

Pavela, G., Y.I. Kim, and S.J. Salvy. 2018. "Additive Effects of Obesity and Loneliness on C-reactive Protein." *PLOS One* 13(11): e0206092. DOI: https://dx.doi.org/10.1371/-journal.pone.0206092.

Peytchev, A. and E. Peytcheva. 2017. "Reduction of Measurement Error Due to Survey Length: Evaluation of the Split Questionnaire Design Approach." *Survey Research Methods* 11(4): 361–368. DOI: http://dx.doi.org/10.18148/srm/2017.v11i4.7145.

Raghunathan, T.E. and J.E. Grizzle. 1995. "A Split Questionnaire Survey Design." *Journal of the American Statistical Association Statistical Association* 90(429): 54–63. DOI: https://dx.doi.org/10.2307/2291129.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberg. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27(1): 85–95.

Raghunathan, T.E., P.W. Solenberger, and J. van Hoewyk. 2002. *IVEware: Imputation and Variance Estimation Software User Guide*. Ann Arbor, MI: Institute for Social Research, University of Michigan. Avalilable at: ftp.isr.umich.edu/pub/src/smp/ive/i-ve_user.pdf (accessed September 2014).

Rhemtulla, M., and T. Little. 2012. "Tools of the Trade: Planned Missing Data Designs for Research in Cognitive Development." *Journal of Cognition and Development: Official Journal of the Cognitive Development Society* 13(4). DOI: https://dx.doi.org/ 10.1080/ 15248372.2012.717340.

Roszkowski, M.J., and A.G. Bean. 1990. "Believe it or not! Longer Questionnaires have Lower Response Rates." *Journal of Business and Psychology* 4(4): 495–509. DOI: https://dx.doi.org/10.1007/BF01013611.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63(3): 581–592. DOI: https:// dx.doi.org/10.2307/2335739.

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

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

Shah, M., D. Paulson, and V. Nguyen. 2018. "Alcohol Use and Frailty Risk among Older Adults over 12 Years: The Health and Retirement Study." *Clinical Gerontologist* 41(4): 315–325. DOI: https://dx.doi.org/10.1080/07317115.2017.1364681.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 47(1): 36–53. DOI: https://dx.doi.org/10.1086/268765.

Shoemaker, D.M. 1973. *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger Publishing Company.

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

Thomas, N., T.E. Raghunathan, N. Schenker, M.J. Katzoff, and C.L. Johnson. 2006. “An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey.” *Survey Methodology* 32(2): 217–231.

Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. “MICE: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software* 45(3): 1–67. Doi https://dx.doi.org/10.18637/jss.v045.i03.

Wagner, J., and K. Olson. 2018. “An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys.” *Journal of Official Statistics* 34: 211–237. DOI: https://dx.doi.org/10.1515/jos-2018-0010.

Yansaneh, I.S. 2005. “An Analysis of Cost Issues for Surveys in Developing and Transition Countries.” In *Household Sample Surveys in Developing and Transition Countries*. 253-266. New York: United Nations. Available at: https://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf (accessed November 2019).

Zabel, J.E. 1998. “An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior.” *The Journal of Human Resources* 33(2): 479–506. DOI: https://dx.doi.org/10.2307/146438.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016. “Synthetic Multiple-Imputation Procedure for Multistage Complex Samples.” *J Off Stat* 32(1): 231–256. DOI: https://dx.doi.org/10.1515/JOS-2016-0011.

Received October 2018
Revised March 2019
Accepted April 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 855–886, http://dx.doi.org/10.2478/JOS-2020-0041

Double Barreled Questions: An Analysis of the Similarity of Elements and Effects on Measurement Quality

Natalja Menold[1]

In double barreled questions (DBQs) respondents provide one answer to two questions. Assumptions how respondents treat DBQs and how DBQs impact measurement quality are tested in two randomized experiments. DBQs are compared with revisions in which one stimulus was retained while the other stimulus was skipped. The observed means and parameters when modeling latent variables differed among the versions. Metric and scalar measurement invariance was not given among the versions, and at least one single stimulus version was found to be associated with a higher validity. Response latencies did not differ among versions or respondents needed less time to respond to DBQs. The author concludes that respondents may understand the stimuli in a DBQ differently, and access one of them while disregarding the other, which can have an adverse effect on validity.

Key words: Question wording; double barreled questions; validity; comparability; measurement invariance.

1. Introduction

The recommendation to avoid double-barreled questions (DBQs) has been repeated in various textbooks (e.g., Bradburn et al. 2004; Dillman et al. 2014; Le Payne 1951; Oppenheim 1992) since the earliest days of survey research in the 1940s. DBQs present more than one aspect, such as opinions or behaviors, together in a single question so that "respondents must answer two questions with one answer" (Bradburn et al. 2004, 142).

Nontheless, despite recommendations to the contrary, DBQs remain very common in surveys and inventories. The question in the European Social Survey (ESS 2014) on media use, for example, requires respondents to evaluate two or more objects in a single question: "And again, on an average weekday, how much of your time watching television is spent watching news or programs about politics and current affairs?". Respondents have to estimate the amount of time they spend watching programs about politics – the first stimulus – and current affairs – the second stimulus. Further examples can be found in established inventories. The PVQ, Portrait Values Questionnaire (Schwartz 2003, 286), which is also included in the ESS and other large-scale surveys (e.g., GESIS-Panel), consists of questions containing at least two stimuli, each as a sentence: "How much are you like this person? He/she believes that people should do what they're told. He/she thinks people should follow rules at all times, even when no-one is watching."

[1] Institute of Sociology, Technische Universität Dresden, D-01062 Dresden, Germany. Email: natalja.menold@tu-dresden.de

Including more than one stimulus in a single question increases the question's complexity, which the survey research literature (e.g., Dillman et al. 2014) demonstrates is an undesirable property. Complexity may be associated with ambiguous and complex linguistic and grammar structures, such as the use of vague and imprecise terms (Graesser 2006; Lenzner et al. 2010), by the number of clauses (Yan and Tourangeau 2008), but also solely by the number of words and the number of syllables (Le Payne 1951). The complexity of survey questions, of which DBQs are a special instance, might increase cognitive burden for respondents so that they also take longer to respond to them (Lenzner et al. 2010). However, as Lenzner et al. (2010) discuss, we understand very little about what makes a question difficult or what imposes a cognitive burden. Likewise, although many authors expect DBQs to be associated with high cognitive burden and lower data quality (Dillman et al. 2014; Le Payne 1951; Oppenheim 1992), little is known about how DBQs impact the cognitive response process and measurement quality. With a focus on measurement of opinions in multi-item inventories, the present article describes experimental studies that investigate (1) how respondents respond to the stimuli in DBQs and (2) the impact of DBQs on response time and the quality of measurement of latent variables. The author formulates research hypotheses in the next two sections based upon the discussion of complexity of survey requests and respondents' cognitive process associated with DBQs. The author goes to describe the method and data of experimental studies and their results. Finally, discussion of the results and implications from the study are provided.

1.1. Complexity Associated With DBQs

The complexity of DBQs concerns two aspects: i) the wording of a question and ii) the cognitive process, which the complexity elicits in respondents. With respect to question wording, DBQs are linguistically more complex than analogous questions that contain just single stimuli (Single Stimulus Questions, SSQs). DBQs that form part of an independent sentence or a dependent clause in a complex sentence are themselves less complex. In such sentences, DBQs have every grammatical function that a word takes in a sentence: double subjects, activities, objects or those attributes. More complex, however, is the case of DBQs that consist of two (or even more) single sentences (sentences which may themselves be simple or complex), such as those of PVQ.

How does the cognitive process involved in responding to numerous stimuli differ from the process of responding to questions with one single stimulus? In contrast to SSQs, DBQs require an evaluative single response to two (or more) potentially different aspects (Oppenheim 1992; Le Payne 1951). This means that the cognitive process involved in responding to a DBQ, described by Tourangeau et al. (2000) as consisting of steps (1) comprehension, (2) retrieval, (3) judgement and (4) response, differs from the process of responding to a complex question that is not a DBQ. If a question is complex, but does not contain two or more stimuli for evaluation, respondents have difficulties at the comprehension stage in particular. However, in the case of a DBQ, if we assume a complex SSQ and a DBQ to be comparable in terms of complexity, respondents have to complete an additional task at the stage of comprehension, namely evaluate whether the two stimuli are similar or different in meaning. In the case of a complex SSQ, complexity may not itself effect retrieval and judgement. Once respondents have understood the question (although not without difficulties), they can retrieve and use memory information to respond to the question in the same way as they would to an easy question. Retrieval and judgement are more complex for DBQs than for SSQs, as for DBQs, respondents have to remember information und use it to form a response to each of the stimuli. If respondents evaluate the stimuli as similar, they can then arrive at a compound response for both stimuli. If they evaluate the stimuli to be different in meaning, it is more difficult or even impossible to generate a compound of such two stimuli, which reflects the high level of complexity of the respondents' cognitive task.

Some authors, for example, Bradburn et al. (2004, 142), discuss questions of the type "Are you in favor of building more nuclear power plants so that we can have enough electricity to meet the country's needs, or are you opposed to more nuclear power plants even though this would mean less electricity", as well as the attribution of attitudes to well-known persons as double-barreled. However, other authors do not classify such questions as DBQs, but as "presuppositions" (Kay and Fillmore 1999; Dillman et al. 2014). Although presuppositions consist of different parts, like DBQs, they are distinct from DBQs in an important way. DBQs are questions that contain either specifications or features of one object, subject, or activity, or are grammatically independent enumerations of distinct objects, subjects, activities, and so on. There is a relationship of grammatical dependence between stimuli in a presupposition, as the stimuli either present a source and its effect, for example, or one barrel is object, subject, or activity while the other represents an attribute given to it. In a DBQ, each aspect should be responded to. By way of contrast, one of the barrels in a presupposition is the subject of evaluation, whilst the other represents facts or beliefs that are taken for granted (Kay and Fillmore 1999). The above example by Bradburn et al. (2004) takes it as given that more electricity can be generated with nuclear power plants (and in addition that this will enable the country to meet its needs). Respondents are assumed to be in agreement with premises that may in fact be wrong. The cognitive process for dealing with a presupposition would be even more complex than for dealing with a DBQ. When dealing with a presupposition, respondents do not evaluate the similarity or difference of stimuli at the comprehension stage, but have to decide about their agreement with the granted part at the judgement stage. Dillman et al. (2014) classify presuppositions differently from DBQs as questions that do not apply to every respondent. Similarly, the present article understands DBQs to be distinct from presuppositions. The author also avoids mixing questions with multiple stimuli that are connected with "and" and "or". Although questions with "or" could be seen as a special form of DBQs, the cognitive process may be different because respondents have to select a stimulus and can disregard the other. In the case of "and" and similar DBQ constructions, such a choice is not offered from the outset. Therefore, the present research does not address both presuppositions and double stimuli connected with "or".

If the task of respondents is difficult and their motivation or cognitive abilities are low, respondents might lean to satisficing behavior (e.g., Krosnick 1991; Krosnick and Presser 2009). Satisficing means that respondents who decide that the stimuli in a DBQ are different may just focus on one of them and disregard the other. It is also likely that respondents would even stop evaluating the similarity of two stimuli and pay attention to only one of them.

As mentioned before, the common argument against DBQs is that respondents would be too confused to provide a response to two parts with different meanings. However, as empirical research on DBQs is rare, the cognitive process or the difficulties that respondents might have with them are not well understood. For example, we do not know what stimuli are evaluated by respondents as similar or different or in what cases. Research on this question is important because researchers who use DBQs in their inventories will have assumed that they transport similar meaning or that they are two parts of a compound entity (as otherwise they would have avoided posting ambiguous questions). The assumption that different stimuli in a DBQ would complement each other may be erroneous, as a study by Grant Levy (2019) demonstrates. In a randomized experiment with university students, Grant Levy (2019) compared the complex response alternatives (question from a GALLUP survey) with separate responses to each of the stimuli. The respondents' task was to "Circle the number that represents the statement that is closest to your belief" (Grant Levy 2019, 1996). In the first experimental condition, statements such as "God created human beings pretty much in their present form at one time within the last 10,000 years or so" were used. In the other condition, respondents had to evaluate separately: "How long do you believe human beings have been on the planet in approximately their present physical form? (a) About 5000 years ago (b) About 10,000 years ago (c) About 50,000 years ago (d) About 250,000 years ago" and "Do you believe that humans were created by God either through an evolutionary process or otherwise? (a) Yes (b) No". Grant Levy (2019) found very little correspondence between the conditions. This means that only a small proportion of the respondents selected response combinations that resembled the presented double-barreled version, that is (b) from the first question and (a) from the second question. The findings by Grant Levy (2019) therefore support the assumption that respondents can evaluate the stimuli included in a DBQ differently. Vettehen and Van Snippenburg (2002) evaluated questions of the kind "I watch television to keep up with important events". The authors showed that increasing the number of items that describe theoretically distinct motivations to watch television increased the correlation of the battery with the behavior "watching television." Since the common part of the motivation items and the behavior items was "watching television" and not the specific reason for this, the authors concluded that respondents tend to focus on the first part of the question "I watch television" and not on the specific motivation statement. Although the questions evaluated by Vettehen and Van Snippenburg (2002) would not necessarily be DBQs as defined here, the study provides evidence that in a complex question, respondents might consider only one part of a question and disregard the other.

The question arises as to how respondents treat more typical double-barreled opinion questions that were of present concern, such as "The age in which discipline and obedience to authority are some of the most important virtues should be over" (Aichholzer and Zeglovits 2015). Do the respondents agree that both, "discipline" and "obedience to authority", are not important virtues anymore, or would they think that this applies to the "obedience to authority" and not to "discipline"? Likewise, a respondent can feel herself less similar to a person who "believes that people should do what they're told" and more similar to someone who "thinks people should follow rules at all times, even when no-one is watching." If respondents respond comparably to both stimuli, such stimuli are referred to as similar or parallel. If the responses to the stimuli are not comparable, the stimuli are referred to as different and therefore non-parallel. In line with the rarely tested assumption in textbooks that respondents may evaluate the stimuli differently and therefore have difficulties responding to a DBQ (e.g., Oppenheim 1992) and in the light of the findings by Grant Levy (2019), the first hypothesis is stated as follows:

Hypothesis 1(H1): The single stimuli in DBQs have a different meaning to respondents and are therefore not parallel.

The present article focuses on opinion questions in multi-item instruments that have a desirable property to measure a latent concept or variable of interest, for example "Tradition" as human value in the case of PVQ. As compared to the study by Grant Levy (2019), scalar multi-item measures, that is, established inventories, are used. DBQs in scalar questions on opinions can be split into their separate parts and presented to different independent and randomly divided groups of respondents (Bassili and Scott 1996). To evaluate the degree of parallelism between the stimuli of a DBQ, one can compare manifest means between the groups, as well as measurement invariance (Meredith 1993) by means of LVM, Latent Variable Modeling (Raykov and Marcoulides 2011). Different understandings of stimuli in particular can be investigated by means of measurement invariance analysis when randomly divided groups of respondents evaluate them independently from each other (Hox et al. 2015).

1.2. Effects of DBQs

Linguistically difficult questions were found to be associated with longer response times (Lenzner et al. 2010) or with higher context effects, such as effects of response options (Bless et al. 1992; Le Payne 1951). Some indications that DBQs negatively impact the response process, that is, that respondents have difficulties responding to them as outlined above, come from qualitative cognitive pretesting studies (e.g., Lenzner et al. 2017; Yorkston et al. 2008). It is not clear from these studies, however, whether the revised questions (with one stimulus) are easier to respond to and are thus associated with a higher measurement quality. Some inventory construction studies that document the item revision process provide support for the view that removing or revising DBQs can lead to higher measurement quality. These studies show that the measurement quality of forms that included DBQs and other problematic questions (negatives, double negations, etc.) was improved following a revision (e.g., Campbell et al. 2009; Fowler Jr. 1992; Gemenis 2013, Stafford 2011; Williams et al. 2009). However, many additional revisions beyond addressing DBQs were incorporated, so the improved measurement quality could be due to the removal of DBQs or also to other revisions.

Only a few studies focus on the impact of DBQs on data quality. After a systematic literature search, the author was only able to identify two publications by Borgers and Hox (2001) and by Bassili and Scott (1996). In their non-experimental correlative study on children, Borgers and Hox (2001) could not find any effects of DBQs on item nonresponse and suggested that experimental research on this topic should be conducted. However, in this article it remains unclear what kind of questions are classified as DBQs. Bassili and Scott (1996) compared DBQs with two versions in which the DBQs were split into separate parts using a telephone survey on a students' sample. (Each barrel in the study also contained presuppositions, but the authors split the barrels so that an SSQ was also a presupposition). Respondents experienced more difficulties with the DBQs than with the separated parts because the response latencies and the number of clarifications that respondents requested from an interviewer were higher for the original DBQ versions than for their parts, evaluated separately.

In line with the findings by Lenzner et al. (2010) and Bassili and Scott (1996), the author of the present study expected respondents' cognitive process to be of a greater difficulty in the case of inventories with DBQs than in the case of SSQs. Respondents would consequently need more time to respond to the DBQs than to the analogous questions with single stimuli. This issue is addressed by Hypothesis 2.

Hypothesis 2 (H2): Response times are higher for inventories that include DBQs than for those with single stimuli.

In light of the above discussion about the potential satisficing in response to high-task difficulty, however, higher response times in the case of DBQs are plausible only if respondents carefully process information and spend effort in comprehension, retrieval and judgement.

With respect to complex and ambiguous questions, researchers report negative effects on measurement quality in terms of item nonresponse or evaluations of the effects of questionnaire properties, that is, rating scales or other context effects (e.g., Schaeffer and Dykema 2011). However, previous research has seldom addressed more direct metrics of measurement quality, such as reliability and validity (Schaeffer and Dykema 2011). The author assumes that the effect of DBQs would depend on the parallelism of stimuli in a DBQ with respect to their meaning and use. If stimuli in inventories with DBQs are parallel in meaning, there would not be differences between corresponding inventories with SSQs with respect to the results they provide, which also implies that there would be no differences between them in measurement properties. Therefore, in such a case, inventories with DBQs would not have lower reliabilities and validities than those with SSQs. However, if respondents evaluate the stimuli in DBQs as different and thus as not parallel, this would have a negative effect on data quality, i.e., on non-systematic measurement error or reliability as well as on systematic measurement error or validity (Groves et al. 2009). For the present study, a negative impact of DBQs on measurement quality is assumed as follows.

Hypothesis 3 (H3): Measurement quality (reliability and validity) is higher in inventories with SSQs than those with DBQs.

2. Method and Data

2.1. Instruments

The author conducted two independent, between-group-design experiments using two established inventories that employed DBQs: (1) a German language measure of authoritarianism, the Balanced Short Scale of Authoritarian Opinions (B-RWA-6) (Aichholzer and Zeglovits 2015) in the first experiment and (2) two subscales of PVQ (Schwartz 2003) in the second experiment.

The implementation of these instruments is of advantage, because they are based on established theories and are well documented with respect to the underlying concepts, measurement structure and measurement quality. This allows theoretically justified and empirically proven assumptions to be made about the properties of measurement (i.e., factorial structure, reliability and validity), as well as comparisons of measurements among different question wording forms.

To compare these original DBQ inventories with the use of a reduced number of stimuli, the author implemented two versions with SSQs that contained either one or the other stimulus. This resulted in three randomized groups in each experiment: (1) DBQ version (DBQ group), and two groups with single stimuli versions (2) SSQ1 group and (3) SSQ2 group.

The B-RWA-6 contains six items measuring three conceptual sub-dimensions of authoritarianism. Authoritarianism means support for authorities, being in favor of sanctions in the case of non-conformity, and rigid orientation towards traditions and established norms (Altemeyer 1981). The three sub-dimensions are "authoritarian submission", "authoritarian aggression" and "authoritarian conventionalism". Each sub-dimension consists of two items, one is positively worded, while the other is reversed (balanced scales).

In DBQ questions, which are of interest in the present article, double stimuli are enumerations in a sentence or grammatically independent sentences. DBQs included in a sentence contain multiple stimuli as enumerations, separated by comma or conjunction "and". A question can be a clause with the structure "subject (a man) + verb (reads) + object (a book)". Double stimuli would be two (or more) subjects, activities and objects. In more complex structures, an attribute can be added to each component: "a young man"; "reads with interest"; "a book about politics". Double stimuli can exist for such attributes as well. Similarly, double stimuli can be included in the subordinate clauses.

In the B-RWA-6 inventory, four items of six are double-barreled (sub-scale A "Submission" and B "Conventionalism"). Table 1 provides an overview of the wording of the items of the B-RWA-6 and their pertinent labels, AUT1 to AUT6. The double stimuli in the B-RWA-6 are components of independent and/or subordinate clauses. For example, let us look at the first item of the "Submission" sub-dimension (AUT1): "We should be grateful for leaders who tell us exactly what we shall do and how." In this item, "What" (we shall do) is the first and "How" is the second stimulus. If we look at the structure of the sentence, double stimuli are components of the subordinate clause and represent an object and an attribute of the verb "shall do". In the item AUT2 double stimuli are two subjects, and in the item AUT5 they are two objects, included in the subordinate clause. In AUT6, there are two verbs that describe the activity of the subject (people) in the subordinate clause as well.

The four items with double stimuli were reworded to obtain a single stimulus version. For example, for the item AUT1, one form (SSQ1 group) contained only the stimulus "What we shall do", while the second stimulus "How" was dropped resulting in the wording "We should be grateful for leaders who tell us exactly what we shall do". The wording of the other counterpart (SSQ2 group) was "We should be grateful for leaders who tell us exactly how we shall do something" (Table 1). To attain to the possible presentation order effects of the double stimuli in the original version, the author implemented a mixed order of the stimuli in the revised versions. So for the items AUT2 and AUT5 (Table 1), the second stimulus of the original DBQ was implemented in the SSQ1 version and the first stimulus of the original DBQ in the SSQ2 version. For the two remaining items (AUT1, AUT6) the SSQ1 version contained the first, and the SSQ2 version contained the second double stimulus. The items of the sub-dimension B "Aggression" (AUT3 and AUT4) did not contain double stimuli, so each experimental group used them without variations. The response formats also did not vary between the experimental groups (see Notes, Table 1).

Table 1. Question wording of the items of the B-RWA-6 used in the experimental groups.

Item	Original DBQ	SSQ1	SSQ2	Subscale
AUT1	We should be grateful for leaders who tell us exactly what we shall do and how.	We should be grateful for leaders who tell us exactly what we shall do.	We should be grateful for leaders who tell us exactly how we shall do something.	A
AUT2	The age in which discipline and obedience to authority are some of the most important virtues should be over.	The age in which obedience to authority is one of the most important virtues should be over.	The age in which discipline is one of the most important virtues should be over.	A
AUT3	Our society for once has to crack down harder on criminals	Our society for once has to crack down harder on criminals	Our society for once has to crack down harder on criminals	B
AUT4	It is important also to protect the rights of criminals	It is important also to protect the rights of criminals	It is important also to protect the rights of criminals	B
AUT5	This country would flourish if young people paid more attention to traditions and values.	This country would flourish if young people paid more attention to values.	This country would flourish if young people paid more attention to traditions.	C
AUT6	Our country needs people who oppose traditions and try out different ideas.	Our country needs people who oppose traditions.	Our country needs people who try out different ideas.	C

Notes.

1) Source of the items in the DBQ version: Aichholzer and Zeglovits (2015).

2) A = Support for Authorities; B = Aggression; C = Conventionalism; The question wording of AUT3 and AUT4 was not varied among the experimental groups, because the items were not DBQs.

3) Instruction (used with each version): "Please provide, how much a statement does apply."

4) Response options (used with each version): does not applies at all (*trifft überhaupt nicht zu*); applies slightly (*trifft wenig zu*); applies to some extent (*trifft einigermaßen zu*); applies to a great extent (*trifft ziemlich zu*); fully apples (*trifft voll und ganz zu*).

The inventories with DBQs like the B-RWA-6 are more typical in questionnaires, as multi-item sets usually consist of single sentences with or without subordinate clauses, where the double stimuli are two or more subjects, verbs, objects or their attributes. The PVQ inventory (Schwartz 2003) has a different structure because *two separate grammatically independent sentences* describe the subject that respondents have to evaluate with respect to the similarity to themselves. Some of the separate sentences are simple, while some contain dependent clauses or even double-barreled objects (e.g., item Trad1). Although this two-sentence structure of the PVQ is a less typical form of DBQ, it is worth comparing it with more typical DBQ structures, such as in B-RWA-6. The potential cognitive burden of DBQs might not only be due to their grammatical complexity, but more particularly to the task of evaluating two fewer or more different aspects as a single entity. Respondents have to conduct this task in the case of both inventories; therefore, the results obtained from them would be similar, although the inventories have a different structure. In addition, PVQ was included, because it is a relevant inventory used in large-scale population surveys, such as the ESS or GESIS Panel.

The author used four items of the PVQ (Schwartz 2003) from the questionnaire of the German ESS (2014). The PVQ measures ten basic values, grouped into four second-order values. The items used in this experiment assess the basic values “Tradition” and “Conformity”, which belong to the second-order value “Conservation.” “Tradition” and “Conformity” are strongly related to each other (latent correlation of .98, reported by Schwartz 2003). Tradition describes respect for and acceptance of traditional cultural or religious customs. Conformity is defined as a restraint of actions, inclinations, and impulses likely to upset or harm others who violate social expectations or norms (Schwartz 2003). The instruction and response alternatives – presented in Table 2 – were those of the ESS (2014) questionnaire. The original items are included in the column “original DBQ” in Table 2. The first and the second sentences of the original items of the PVQ differ in form, as the first sentence is shorter and contains a statement about importance of a value or a general kind of thinking about a value, whereas the second statement is somewhat longer and describes a tendency to a more concrete behavior or opinion. The first revised version therefore contained the first sentence (SSQ1 group), while the other revised version contained the second sentence (SSQ2 group) of the original PVQ items. Because of the differences in the length and structure of the sentences, the author decided to consistently incorporate the first sentence in one form and the second sentence in the other form and not to mix them in the forms. Otherwise, respondents could be confused when reading differently realized sentences in one form that might negatively affect measurement quality.

2.2. *Participants*

The data was collected in December 2016 using a commercial online access panel with a proportional quota sample approximately reassembling the German adult population aged 18 years or older. The participants used only a PC to avoid the side effect of other devices. Device was controlled at the beginning of the session; use other devices than PC lead to the exclusion from the sample.

Of the participants in the first experiment with the B-RWA-6 inventory (N = 497), 50.3% were men. With respect to education, 29.6% had completed senior high school and

Table 2. Question wording of the items of the Schwartz values (PVQ) used in the experimental groups.

Item	Original DBQ	SSQ1	SSQ2	Subscale
Trad1	It is important to him/her to be humble and modest. He/she tries not to draw attention to himself/herself.	It is important to him/her to be humble and modest.	He/she tries not to draw attention to himself/herself.	Tradition
Trad2	Tradition is important to him/her. He/she tries to follow the customs handed down by his/her religion or his/her family.	Tradition is important to him/her.	He/she tries to follow the customs handed down by his/her religion or his/her family.	Tradition
Conf1	He/she believes that people should do what they're told. He/she thinks people should follow rules at all times, even when no-one is watching.	He/she believes that people should do what they're told.	He/she thinks people should follow rules at all times, even when no-one is watching.	Conformity
Conf2	It is important to him/her always to behave properly. He/she wants to avoid doing anything people would say is wrong.	It is important to him/her always to behave properly.	He/she wants to avoid doing anything people would say is wrong.	Conformity

Notes.

1) Source for translated items: Schwartz (2003); Source for the items used in the experiment: German ESS questionnaire.

2) Instruction: "How much is the described person like or not like you."

3) Response Options (used in each experimental group): very similar (*ist mir sehr ähnlich*), similar (*ist mir ähnlich*), somewhat similar (*ist mir etwas ähnlich*), a little bit similar (*ist mir nur ein kleines bisschen ähnlich*), not similar (*ist mir nicht ähnlich*), not similar at all (*ist mir überhaupt nicht ähnlich*); do not know (*weiß nicht*)

30.8% secondary school. Concerning the age, 20% were younger than 40, 34.6% were between 40 and 60 and 35.4% were 60 and older.

Of the participants in the second experiment with the PVQ inventory (N = 435), 47.8% were men; 29.4% had completed senior high school and 30.6% had completed secondary school; 26.2% were younger than 40, 38% were between 40 and 60, and 35.9% were 60 and older.

In both experiments, none of the experimental groups differed significantly ($p > .10$) with regard to these respondent variables; the pertinent hypothesis was tested with a χ^2 test.

2.3. Data Analysis

To address hypotheses H1 and H2 mean differences among versions were compared by means of Multivariate Analysis of Covariance (MANCOVA) and Univariate Analysis of Covariance (ANCOVA) with the SPSS 23 software. To control for the possible effects of respondents' gender, education, and age, the author included these variables in the analysis as auxiliary variables (covariates). To compare structures of inventories with respect to the latent variable measurement, the author made use of the framework of LVM (Muthén 2002) and implemented Multi Group Confirmatory Factor Analysis (MGCFA) with the Mplus 8.2 software. The author thereby examined measurement invariance between different experimental groups. This allows for an evaluation of the parallelisms of the versions with respect to their relevance to the latent dimension and comparability of measurements. The author first tested the exact measurement invariance (Meredith 1993) by evaluating configural, metric and scalar invariance. Configural invariance means that the same latent structure explains the observed means, variation, and covariation among the items in each of the groups, but does not allow for any comparisons between groups (Meredith 1993). Metric invariance means that an indicator has comparable strength of linear relationship with the latent variable and thus is equally relevant to it in each version (equality of factor loadings). Establishing metric invariance allows for comparison of correlations between latent variables or summarized scores (Hox et al. 2015; Meredith 1993) among the respective groups. Scalar invariance implies that the means of each of the observed variables are comparable associated with the latent means and therefore tap similar parts of latent continuum (equality of intercepts). Scalar invariance allows for a comparison of latent means or summarized scores (Hox et al. 2015; Meredith 1993).

The model fit of MGCFAs was evaluated using the chi-square (χ^2) test, the Root-Mean-Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI) (Beauducel and Wittmann 2005). The CFI should be 0.95 or higher, while an RMSEA of 0.08 or less indicates an acceptable fit (Hu and Bentler 1999). Robust Maximum Likelihood estimator (MLR) was used due to the non-normality of data in each experiment (Muthén and Muthén 2017). Concerning the exact measurement invariance, a significant change of χ^2 (Meredith 1993) or a change of ΔCFI $\geq .01$ and ΔRMSEA $\geq .015$ indicate significant differences (Chen 2007).

In the structural equation modeling literature (Muthén and Asparouhov 2014), there is some criticism that the exact test of measurement invariance is too conservative and that metric and scalar invariance are consequently difficult to reach. The author therefore also implemented a more liberal alignment method (Muthén and Asparouhov 2014) using Mplus 8.2 software.

Reliability and validity were evaluated to assess measurement quality and to address the hypothesis H3. While reliability describes how far the variation in the data is due to the true variation and whether the non-systematic measurement error is negligible (Lord and Novick 1968; Raykov and Marcoulides 2011), validity refers to the empirical evidence that an inventory allows making conclusions with respect to the concept under investigation (Kane 2013; Messick 1989).

For the B-RWA-6, Aichholzer and Zeglovits (2015) report the composite reliability ranging between .60 and .65 that is a low, just acceptable size. For the PVQ subscales, Schwartz et al. (2015) report Cronbach's Alpha of a very low size (Alpha = .33 for Tradition and Alpha = .53 for Conformity) they found in the ESS 2012 data in Germany. Therefore, there would be a potential to increase measurement quality through an elimination of DBQs in both inventories.

The author used *Composite Reliability* as a method to assess reliability of different versions. This method is based on the "congeneric measurement model" and does not assume equality of factor loadings or error term covariances associated with the observed measures. The *Composite Reliability* coefficient ($\hat{\rho}_x$) is estimated within the framework of LVM as follows (Raykov and Marcoulides 2011, 161):

$$\hat{\rho}_x = \frac{(\hat{b}_1 + \cdots . + \hat{b}_p)^2}{(\hat{b}_1 + \cdots . + \hat{b}_p)^2 + \hat{\theta}_1 + \cdots + \hat{\theta}_p}, \tag{1}$$

where $b_1, \ldots, b_p$ are the factor loadings and $\theta_1, \ldots, \theta_p$ the error variances, obtained from the MGCFA (see Menold and Raykov 2015).

To assess validity coefficients, the author evaluated convergent (nomological) validity, as well as criterion validity. The convergent validity is evidenced through high correlations between different measures of the same or very closely related concepts (e.g., Lord and Novick 1968). Respondents therefore also administered the Short Scale of Authoritarianism (KSA-3) by Beierlein et al. (2014) as a to the B-RWA-6 alternative inventory on authoritarianism to provide data for the analysis of validity. KSA-3 consists of two subscales "Convention" and "Aggression" (with three items each). In order to evaluate validity, the size of expected correlations on the basis of theoretical considerations and/or results of past research must also be stipulated (e.g., Lord and Novick 1968). KSA-3 and B-RWA-6 represent alternative measurement instruments for the same latent concept (authoritarianism). Therefore, one can expect correlations with high effect sizes (larger than .50, Cohen 1992) between them. Expected correlations are shown in Table 7 in the results section, along with the results with respect to the validity coefficients in different questions wording versions.

The author also evaluated convergent validity for "Conformity" and "Tradition" of the PVQ. For this purpose, respondents administered the subscale "Security" of the PVQ (Schwartz 2003). "Security" can serve as measure of a similar concept, because it is the third sub-dimension of the second-order "Conservation" value, while "Tradition" and "Conformity" are the first two. The latent correlation between "Security" and "Tradition" vs. "Conformity" amounts to .78 in Germany in the ESS 2012 data (reported by Schwartz et al. 2015), so this size of correlation was also expected for the present study (Table 7).

Criterion validity is evidenced through replications of empirically proven relationships between a measure and a third variable (Raykov and Marcoulides 2011). The measures on "Security" of the PVQ can therefore serve as criterion for the B-RWA-6, since Beierlein et al. (2014) found the authoritarianism measured by the KSA-3 to correlate by .65 with "Security". Similarly, the KSA-3 inventory can be a measure of criterion for the PVQ values "Tradition" and "Conformity". Beierlein et al. (2014) reported correlations of the KSA-3 with Schwartz's "Tradition" ($r = .49$) and "Conformity" ($r = .58$), so that comparable sizes of correlation were expected to support validity assumptions in the present study (Table 7).

The author compared latent convergent or criterion validity coefficients among different question-wording groups using MGCFAs and Mplus 8.2 software. A similar procedure is described in detail in Raykov et al. (2018). The latent covariance between the related concepts thereby was divided by the square root of the product of the error variances of the two latent variables.

3. Results

3.1. Parallelism of the Stimuli in DBQs

3.1.1. Mean Differences

The first hypothesis assumes that the stimuli included in a DBQ are different in meaning and therefore not parallel. In such a case, if the DBQ version and its stimuli are randomly distributed among groups, and respondents separately evaluate each, there should be a notable mean difference among these groups. Concerning the items of the B-RWA-6, the ANCOVAs, which were conducted within a MANCOVA, revealed no significant mean differences between the three experimental groups for the first item (AUT1): "We should be grateful for leaders who tell us exactly what we shall do and how" (Table 3). Hence, the stimuli "what we shall do" and "how we shall do" included in the item AUT1 seem to have similar meaning (within the given sentence or context). As to expectations, there were no mean differences among the groups for the items AUT3 and AUT4 either. These items were not DBQs and those presentations did not differ among groups. However, such a result did not emerge for the remaining three items with experimental variation in question wording (AUT2, AUT5, AUT6). Table 3 shows that the means of these items differed significantly among the versions. The pairwise comparisons showed for the item AUT2 that only the SSQ1 group (with the stimulus "obedience to authority" as an important virtue) differed from the DBQ and the SSQ2 groups. The means in the latter two groups did not significantly differ from each other. The SSQ2 group included the term "discipline" as an important virtue, which was also the firstly presented stimulus in the DBQ group. Due to significant mean difference between the SSQ1 and SSQ2 groups, respondents differently evaluate the meaning of "discipline" and "obedience to authority", a result that supports the expectations of the hypothesis H1.

Table 3. B-RWA-6: Results of the ANCOVAs and pairwise comparisons for mean differences between the experimental groups.

	DBQ		SSQ1		SSQ2		*F (2, 491)*	*part.* η^2
Items	Mean	SD	Mean	SD	Mean	SD		
AUT1	2.42	1.02	2.37	0.86	2.28	0.97	1.12	.01
AUT2	3.08	1.10	3.69[a]	0.93	3.02	1.13	19.44***	.08
Pairwise			DBQ – SSQ1 = -0.61***					
			DBQ – SSQ2 = 0.06					
Differences[b]			SSQ1 – SSQ2 = -0.66***					
AUT3	4.20	0.95	.14.11	0.96	4.00	1.08	1.07	.00
AUT4	2.40	1.06	2.47	1.08	2.48	1.12	0.15	.00
AUT5	3.43	1.02	3.60	0.95	2.97[a]	1.05	17.56***	.07
Pairwise			DBQ – SSQ1 = -0.18					
			DBQ – SSQ2 = 0.46***					
Differences[b]			SSQ1 – SSQ2 = 0.63***					
AUT6	2.79[a]	1.03	2.30[a]	0.95	3.83[a]	0.90	109.74***	.30
Pairwise			DBQ – SSQ1 = 0.50***					
			DBQ – SSQ2 = -1.01***					
Differences[b]			SSQ1 – SSQ2 = -1.51***					
n	165		153		179			

Notes. ***$p < .001$;

MANCOVA results with respect to the main effects of (1) question wording: *Wilks-Lambda* = 0.59; $F_{(12,972)} = 24.80$; $p = 0.000$; $\eta^2 = 0.23$; (2) respondents' gender (covariate): *Wilks-Lambda* = 0.97; $F_{(6,486)} = 2.82$; $p = 0.01$; $\eta^2 = 0.03$; (3) respondents' age (covariate): *Wilks-Lambda* = 0.92; $F_{(6,486)} = 7.15$; $p = 0.000$; $\eta^2 = 0.08$; (4) respondents' education (covariate): *Wilks-Lambda* = 0.97; $F_{(6,486)} = 2.71$; $p = 0.01$; $\eta^2 = 0.03$.

[a] the mean of the group differs from the other two.

[b] Bonferroni corrected

The pairwise differences for the item AUT5 also showed a significant mean difference between the SSQ2 group and other two groups, whereas there was no significant mean difference between the SSQ1 group and the DBQ group. The SSQ1 group contained the term "values" as an important issue for consideration by young generations, which is the second stimulus included in the DBQ. This means again that the DBQ and the SSQ1 versions are parallel in the meaning, but there is no parallelism for these both to the SSQ2 version. It also seems that, in contrast to the item AUT2, respondents in the DBQ of the AUT5 item considered the second stimulus presented, "values", when responding to it and disregarded the first stimulus presented. For the AUT6, however, one can see that the three groups were not parallel in meaning, because there were significant mean differences among them. Thus, respondents seemed to consider both stimuli when responding to the DBQ in AUT6, which explains its different meaning, as compared with the both single stimuli. Like the item AUT2, the results for AUT5 and AUT6 therefore supported H1 as well.

Table 4 provides an overview of mean differences for each item among the three experimental groups for the Schwartz's PVQ inventory. The means differed significantly among the groups for three out of four items (Trad1, Trad2, Conf2), in line with the expectation of the hypothesis H1. For both of the "Tradition" items, pairwise comparisons (Table 2) showed that respondents evaluated both parts of the DBQs differently. For these items, respondents also tended to focus on one of the stimuli, and not on both, since there were significant mean differences between the DBQ and only one of the single stimulus groups. For the first item (Trad1), there was no significant mean difference between the DBQ group and the SSQ1 group, which evaluated the first sentence of the DBQ ("important to be humble and modest"). By contrast, the mean of the DBQ, as well as that of the SSQ1 group, was significantly different from the mean of the SSQ2 group with the second sentence of the original DBQ item ("tries not to draw attention to himself/herself"). For the Trad2 item, the SSQ1 that contained the first sentence of the initial DBQ ("tradition is important to her/him") differed from it, while there was no difference in means between the DBQ and the SSQ2 that contained the second statement ("tries to follow the customs"). Here, similar to the item AUT5, in the case of DBQ, respondents seemed to respond to the second stimulus presented and not to the first stimulus. In the case of Conf2 item, like in the case of Trad2, the parts of the DBQ were differently evaluated, and the DBQ differed in meaning from the version containing the first sentence (SSQ1), but not from the version that contained the second sentence (SSQ2).

In sum, the majority of the results supported the assumption in hypothesis H1 that stimuli in a DBQ differ in meaning. The results show that respondents tended to evaluate the stimuli included in a DBQ differently. Respondents also seemed to pursue two different strategies in handling the stimuli in a DBQ: (1) evaluating one of the stimuli or (2) considering both of them. The respondents seemed to apply the first strategy more often, which was the case in two of the four DBQs in the B-RWA-6 inventory and three of the four items in the Schwartz's sub-scales. If respondents applied the strategy to evaluate only one of the stimuli, they did not behave in a systematic manner as they sometimes considered the first stimulus of the DBQ and sometimes only the second one.

Table 4. PVQ Items: Results of the ANCOVAs for mean differences and pairwise comparison between the experimental groups.

	DBQ		SSQ1		SSQ2		*F (2, 429)*	*part.* η^2
Items	Mean	SD	Mean	SD	Mean	SD		
Trad1	4.01	1.50	4.35	1.25	3.28[a]	1.45	22.30***	.09
Pairwise[b]			DBQ – SSQ1 = -0.34					
			DBQ – SSQ2 = 0.73***					
Differences			SSQ1 – SSQ2 = 1.07***					
Trad2	3.90	1.34	2.89[a]	1.40	4.09	1.43	32.71***	.13
Pairwise			DBQ – SSQ1 = 1.01***					
			DBQ – SSQ2 = -0.19					
Differences			SSQ1 – SSQ2 = -1.20***					
Conf1	3.74	1.47	3.42	1.48	3.44	1.47	1.91	.01
Conf2	3.72	1.55	3.15[a]	1.53	3.64	1.56	6.65**	.03
Pairwise			DBQ – SSQ1 = 0.60***					
			DBQ – SSQ2 = -0.11					
Differences			SSQ1 – SSQ2 = -0.50**					
n	134		163		138			

Notes. ***$p < .001$; MANCOVA results with respect to the main effects of (1) question wording: *Wilks-Lambda* = 0.68; $F_{(8,852)} = 22.80$; $p = 0.000$; $\eta^2 = 0.18$; (2) respondents' gender (covariate): *Wilks-Lambda* = 0.99; $F_{(4,426)} = 0.80$; $p = 0.53$; $\eta^2 = 0.01$; (3) respondents' age (covariate): *Wilks-Lambda* = 0.97; $F_{(4,426)} = 3.56$; $p = 0.01$; $\eta^2 = 0.03$; (4) respondents' education (covariate): *Wilks-Lambda* = 1.00; $F_{(4,426)} = 0.41$; $p = 0.80$; $\eta^2 = 0.00$.

[a] the mean of the group differs from the other two.

[b] Bonferroni corrected

3.1.2. Measurement Invariance

In this section, the author evaluates the assumption of the parallelism of the latent structures of different question wording versions by means of measurement invariance analysis. In the first step, the measurement model without any restrictions on equivalence of model parameters (a baseline congeneric model, e.g., Raykov and Marcoulides 2011) was specified. For the B-RWA-6 Aichholzer and Zeglovits (2015) postulated a general factor for authoritarianism behind the three sub-factors, consisting of two items each (Figure 1). Because of the reversed formulation of the items for each sub-factor, Aichholzer and Zeglovits (2015) also modeled a latent "method" variable that claimed for potential acquiescence bias (the "bi-factor model", Reise 2012). The author was not able to use the general factor model with a method effect (Figure 1), because of serious specification problems in the present data (convergence problems). Since the three sub-factors are supposed to represent a latent dimension, one factor can also explain the common variation of the six items. The single-factor MGCFA with a latent method effect was identified, but there were still specification problems (negative error variance of the item AUT1, and three to four of the six items had not-significant factor loadings in all groups). The model fit was non-satisfactory as well ($\chi^2_{(df=26)} = 46.76, p < .001$; *RMSEA* = 0.07; *CFI* = 0.87). The next step, after inspecting modification indexes, was to re-specify the method factor. The author allowed only the items with the positive association with the latent variable to load on the method effect as shown in Figure 2. The method effect was assumed to be similar in each group. For the third group, however, a different method effect was specified for the item AUT2 to correct for the corresponding misspecification. Next, in the DBQ group, a correlated error term between the items AUT1 and AUT2 was implemented. This model yielded an acceptable model fit ($\chi^2_{(df=26)} = 37.39$, $p > .05$; *RMSEA* = 0.05; *CFI* = 0.93). This model also resolved the former problem of negative error variances and non-significant item loadings. The item

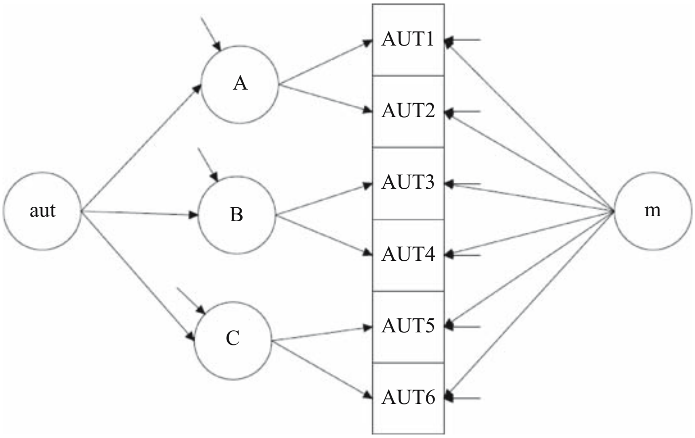


Fig. 1. B-RWA-6 General factor model for the B-RWA-6 with method effect, after Aichholzer and Zeglovits (2015).

Notes. aut: general factor authoritarianism; m: method effect; loadings on m = 1: AUT1, AUT3, AUT5; loadings on m = -1: AUT2, AUT4, AUT6. A = Support for Authorities; B = Aggression; C = Conventionalism; AUT1 TO AUT6 = indicators, see Table 1.

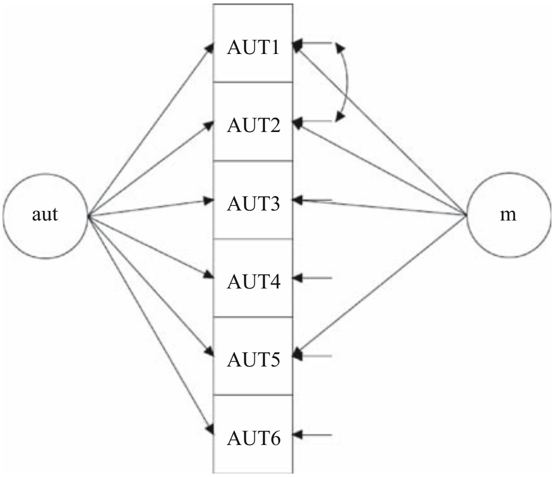


Fig. 2. One-factor model with modified method-effect for the B-RWA-6.
Notes. aut: factor authoritarianism; m: method effect; AUT1 TO AUT6 = indicators, see Table 1. Loadings on m = 1: AUT1, AUT3, AUT5; loading on m = 0: AUT2

AUT6 in the SSQ2 group only did not exhibit a significant factor loading. The described configural model served as a baseline model for the evaluation of measurement invariance, reliability and validity.

The results of the exact measurement invariance test for the B-RWA-6 are provided in Table 5. Restricting the factor loadings being equal, significantly decreased the model fit as accounted by the difference in χ^2, *RMSEA* and *CFI*. There was therefore no exact metric invariance between the groups. Restricting intercepts to being equal further significantly (and dramatically) decreased model fit, meaning that there were notable differences in intercepts between the experimental groups.

The alignment analysis revealed that the factor loading of the AUT2 in the SSQ1 group significantly differed from those in other groups (see Appendix, Section 5). The difference of intercepts between this group and the other two groups are also obtained for AUT2 and AUT5 items. In addition, the intercepts of the item AUT6 differed between the three

Table 5. Results for testing of the exact measurement invariance in two experiments by question wording groups.

Model	$\chi^2(df)$	$\Delta\chi^2(df)$	*RMSEA*	*ΔRMSEA*	*CFI*	*ΔCFI*
			B-RWA-6			
configural	32.39 (26)	-	.051	-	0.928	-
metric	82.37*** (38)	44.98*** (14)	.084	.033	0.718	.210
scalar	281.39*** (45)	208.04*** (8)	.180	.096	0.000	.718
			PVQ			
configural	1.39 (3)	-	0.000	-	1.000	-
metric	24.91*** (11)	24.61*** (8)	0.093	.093	0.934	.066
scalar	182.21*** (19)	171.96 *** (8)	0.240	.147	0.223	.711

Notes. $\Delta\chi^2$: with scaling correction factor for MLR; ****p* < *.001* groups (see Appendix). These results therefore reassembled the results with respect to the differences of means of single items presented in the above section.

To conclude, the latent measurement structure of the B-RWA-6 was different between the two single stimulus groups concerning the loadings and intercepts, a robust result obtained by different methods of testing measurement invariance. The results therefore supported the hypothesis H1 and show that respondents evaluated the stimuli of the DBQs differently so that these stimuli cannot be assumed to be parallel in meaning. The aligment method showed that there were differences between the DBQ version and one SSQ version, while the structures were similar between the DBQ and the other version with SSQs.

For the “Conformity” and “Tradition” of the PVQ, the configural two-factor model (Schwartz 2003) was associated with a perfect goodness-of-fit ($\chi^2_{(\mathrm{df}=3)} = 1.39$, $p = 0.71$, $RMSEA = .00$, $CFI = 1.00$). The item loadings ranged from .37 to .89. Therefore, the configural invariance can be assumed among the groups (Table 5). Modeling the equality of factor loadings significantly decreased the goodness of fit, according to the change in χ^2, *RMSEA* and *CFI.* Restricting the intercepts to being equal noticeably decreased the model fit even further. The exact metric and scalar invariance were therefore not observed between the DBQ and its single stimulus versions. The results of the test of measurement invariance with the alignment method are shown in Appendix. The SSQ1 group had significantly smaller latent means for both latent factors than the DBQ group. The latent mean for Tradition in the SSQ1 group was also lower than in the SSQ2 group. Factor loadings and intercepts for both two items of “Tradition” differed among the experimental groups; a to the comparison of single items comparable result.

The results for both concepts were therefore comparable, regardless of which measurement invariance test was used, and supported the assumption of the hypothesis H1 concerning the difference of meaning between the stimuli of the DBQs.

3.2. *Response Times*

In each experiment, all the items of a version were presented on a screen and response times per screen in milliseconds were measured (absolute response times). Outliers and skewed response times were dealt with by transforming data as described by Yan and Tourangeau (2008). Firstly, data were recoded separately for each experiment, while observations beyond the upper and lower one percentile were replaced with the upper and lower one percentile values. Secondly, the recoded data were log transformed. Because the length of each text version was different, the author also compared response times in milliseconds per character to disentangle the cognitive effort associated with the response process and the different amounts of time needed to read each length of the text. To calculate response times per character in each experimental group, the recoded response times were divided by the number of characters and then log transformed. Table 6 shows the means and standard deviations for logarithmically transformed data by experimental group, as times per screen and times per character.

For the B-RWA-6, a significant difference in response times per character only existed between the two single stimulus groups (see Table 6, pairwise comparisons), where respondents spent more time processing a character of the SSQ1 version. The DBQ group

Table 6. Response times: Results of the ANCOVAs for mean differences between the experimental groups.

	DBQ		SSQ1		SSQ2		*F (df1, df2)*	*part.* η^2
	Mean	SD	Mean	SD	Mean	SD		
			B-RWA-6					
Log scale ms.	6.35	0.47	6.35	0.49	6.30	0.38	0.61 (2, 494)	.00
Log scale ms./char	1.60	0.48	1.69	0.49	1.56	0.37	3.82* (2, 494)	.02
Pairwise			DBQ – SSQ1 = -0.09					
			DBQ – SSQ2 = 0.04					
Differences[a]			SSQ1 – SSQ2 = 0.13*					
Number of characters	114		104		105			
			PVQ					
Log scale ms.	6.22	0.44	6.29	0.43	6.36	0.40	4.83* (2,464)	.02
Pairwise			DBQ – SSQ1 = -0.08					
			DBQ – SSQ2 = -0.15**					
Differences[a]			SSQ1 – SSQ2 = -0.07					
Log scale ms./char	1.35	0.44	1.89	0.43	1.69	0.40	68.00*** (2,464)	.23
Pairwise			DBQ – SSQ1 = -0.54***					
			DBQ – SSQ2 = -0.35***					
Differences[a]			SSQ1 – SSQ2 = 0.20***					
Number of characters	130		82		107			

Notes. ****p* < .001;**p* < .05

[a] Bonferroni corrected

did not differ from the single stimulus groups. For the Schwartz's subscales, there was considerably more text to read in the case of DBQs; the least amount of reading text was in the SSQ1 version. Surprisingly, there was no difference in response times between the DBQ and SSQ1 versions. Response times were significantly lower in the DBQ than in the SSQ2 version. In addition, respondents spent significantly less time per character on the texts in the DBQ than in the other two groups. Respondents spent most time per character in the SSQ1 group, as compared with the other groups (Table 6).

Since respondents did not spend more time on DBQs than on the versions with single stimuli or they even needed less time to respond to DBQs than to their shorter single stimulus versions, there was no empirical support for hypothesis H2 that expected higher response latencies for DBQs. The results with respect to response times therefore support the interpretation of the findings presented in the above section that respondents tended to disregard a part of the question in the case of DBQs and exhibited satisficing behavior.

3.3. Reliability

Reliability was calculated on the basis of the configural models for both inventories (Table 5). In the first experiment, the reliability of the six items of the B-RWA-6 adhering to a single latent dimension was estimated. The correlated error term in the DBQ group was attained to the error variance (Raykov 2012). The composite reliability was estimated at $\rho = .61$ (SE $= .05$; 95% C.I. [.51-.71]) in the DBQ group, which is a low reliability. The reliability was also low in the SSQ1 ($\rho = .61$; SE $= .04$; 95% C.I. [.53-.70]) and SSQ2 group ($\rho = .54$; SE $= .07$; 95% C.I. [.41-.68]) and there were no significant differences in the size of reliability coefficients between the three versions of the questionnaires.

For the PVQ items, the composite reliability for general structure (Raykov 2012) was again low for all experimental groups. It was somewhat higher for the DBQ questions ($\rho = .65$; SE $= .05$; 95% C.I. [.56-.74]) than for the questions with the first sentence (SSQ1: $\rho = .60$; SE $= .05$; 95% C.I. [.50-.70]) and did not differ appreciably between DBQ questions and questions with the second statement (SSQ2: $\rho = .68$; SE $= .04$; 95% C.I. [.60-.77]). The differences between the groups were again not significant.

In sum, the differences concerning reliability were not significant, so that – when considering reliability – hypothesis H3 that expected increased reliabilities in single stimulus questions has to be rejected.

3.4. Validity

Table 7 provides an overview of differences in validity coefficients among the experimental groups. We look at the correlations between the B-RWA-6 and the KSA-3 (convergent validity for the B-RWA-6) in experiment 1. For the "Conservation" of the KSA-3, there was a correlation with the B-RWA-6 of expected high value (.50 or higher) in the DBQ and SSQ2 group. For the "Aggression" of the KSA-3, the expected high value (.50 or higher) was again reached in the SSQ2 group but not in the SSQ1 and DBQ groups. However, for both KSA-3 scales there were only significant differences in validity coefficients between the two single stimulus groups. In addition, in the SSQ1 group there was no significant relationship between the B-RWA-6 and "Aggression" of the KSA-3, which means that two alternative measures of the same concept did not significantly inter-correlate.

Table 7. Validity coefficients, expected, by question wording group and their pairwise differences.

Concept	DBQ	SSQ 1	SSQ2	Expected
		B-RWA-6		
Conservation (KSA-3)	-.62*** (.25) [-.92 – (-.17)]	-.39*** (.11) [-.62 – (-.29)]	-.66*** (.10) [-.82 – (-.65)]	≥ |.50|
Pairwise differences		D (DBQ – SSQ1) = 0.22 (0.26)		
D (SE)		D (DBQ –SSQ2) = 0.41 (0.27)		
		D (SSQ1 –SSQ2) = 0.26* (0.14)		
Aggression (KSA-3)	-.41* (.19) [-.76 – (-.13)]	-.17 (.15) [-.62 – (-.02)]	-.66*** (.09) [-.81– (-.45)]	≥ |.50|
Pairwise differences		D (DBQ – SSQ1) = 0.24 (0.25)		
D (SE)		D (DBQ – SSQ2) = 0.25 (0.21)		
		D (SSQ1 – SSQ2) = 0.48*** (0.17)		
Value security	.28 (.15) [-.02 – .58]	.37**(.12) [.14 – .60]	.63*** (.12) [.39 – .87]	~ .65[1]
Pairwise differences		D (DBQ – SSQ1) = 0.09 (0.20)		
D (SE)		D (DBQ – SSQ2) = 0.35* (0.19)		
		D (SSQ1 – SSQ2) = 0.26* (0.16)		
		PVQ		
Conservation (KSA-3)	-.49*** (.12) [-.71– (-.27)]	-.57*** (.11) [-.76 – (-.35)]	-.18 (.13) [-.55 – (- .04)]	~ |.50-.60|[1]
Pairwise differences		D (DBQ – SSQ1) = .08 (.15)		
D (SE)		D (DBQ – SSQ2) = .33* (.17)		
		D (SSQ1 – SSQ2) = .41* (.17)		
Aggression (KSA-3)	-.34*** (.11) [-.57 – (-.16)]	-.48*** (.10) [-.70 – (-.29)]	-.10 (.13) [-.36 – .15]	~ |.50-.60|[1]
Pairwise differences		D (DBQ – SSQ1) = .14 (.14)		
D (SE)		D (DBQ – SSQ2) = .25 (.16)		
		D (SSQ1 – SSQ2) = .38* (.16)		
Value security	.55*** (.09) [.38 – .72]	0.46*** (.11) [.25 – .68]	0.43*** (.12) [.22– .64]	~ .78[2]
Pairwise differences		D (DBQ – SSQ1) = .09 (.14)		
D (SE)		D (DBQ – SSQ2) = .12 (.14)		
		D (SSQ1 – SSQ2) = .03 (.16)		

Note. *p < .05; **p < .01; ***p < .001; [1]Source: Beierlein et al. 2014; [2]Source: Schwartz et al. 2015.

For the criterion validity of the B-RWA-6, evaluated through the latent correlation with the Schwartz's value "Security", the lowest and non-significant correlation was in the DBQ group and it was higher in both single stimulus groups. The expected size of correlation was only reached in the SSQ2 group, where it was also significantly higher than in the DBQ (D = .35, p < .05) and the SSQ1 group (D = .26, p < .05).

In the case of the PVQ, the latent correlations of "Conformity" and "Tradition" with the subscale "Conservation" of the KSA-3 were somewhat higher in the SSQ1 group than in the DBQ group, but these correlations were within the expected range (.50–.60). The result in the SSQ2 group was critical because a significant latent correlation with the

"Conservation" as measured by KSA-3 was not obtained. The pattern for the latent correlations with the "Aggression," measured by the KSA-3, was similar: The correlation in the SSQ1 group was close to the benchmark, which was not the case in the DBQ group, while the correlation was very low and non-significant in the SSQ2 group. One remarkable result was again that there were strong and significant differences between the two single stimulus groups. Thereby, there was a strong validity loss for one of the parts of the former DBQs (the second sentences), while the validity scores tended to be higher than in the DBQs for the other parts (the first sentences). Inspection of the correlations with a related value (Security) showed higher values in the DBQ than in both single stimulus groups. However, none of the correlations met the benchmark and the differences among groups are neither highly pronounced, nor significant.

The findings of two experiments were therefore comparable. The convergent or criterion validity coefficients were either equal or higher in one of the single stimuli groups than in the DBQ group, but were lower in other single stimuli group. The results provided some support for hypothesis H3. Using SSQs instead of DBQs could increase validity, but this only applied to one of the SSQ versions. The evidence for validity was insufficient with the other SSQ version. This also showed that the respondents considered or evaluated the two stimuli in DBQs differently.

4. Discussion and Conclusions

One of the research aims of the present experimental study was to show that the double stimuli in standard social science inventories could be evaluated as having different meanings (stipulated in hypothesis H1). This assumption is found in many textbooks that discuss why DBQs would be problematic (Le Payne 1951; Oppenheim 1992). However, because there is a lack of empirical evidence on DBQs and because they are very common in inventories, the author considered that those who create inventories might not be aware of potential serious differences of stimuli and assume that they transport similar or comparable meaning and suit to each other. The results for both inventories under investigation mainly support the hypothesis H1 and show that stimuli included in a DBQ have a different meaning for the respondents. The author evaluated eight items (from two inventories) and for six of them significant mean differences were obtained, if the stimuli were individually responded to. This clearly supports the assumption the meaning of the stimuli may potentially differ in a majority of the material of the present study. Looking at the item sets as instruments to measure latent variables, lack of metric and scalar measurement invariance among the versions clearly supports the assumption of difference in the meaning of the stimuli. Absence of metric invariance means that the stimuli differ with regard to their relevance to the latent variable (strength of the linear relationship with the latent variable). Lack of scalar invariance means that stimuli are tapping different latent means and therefore differ not only in the strength of the relationship with the latent variable, but also in the meaning ascribed to it.

The author used two established and well-documented inventories on opinions with very differently structured double stimuli to evaluate the assumption of potential differences in meaning between the stimuli of DBQs. Despite these different structures, a common feature of the DBQs in both inventories was such that respondents had to evaluate two issues as part of a compound in one single question. If the results obtained from the two inventories were not comparable, the effect could not be explained by this commonality of the DBQs. Hence, the results for both inventories were strongly comparable and this strengthens the conclusions with respect to this common problem of DBQs.

Differences or similarities in stimuli in a DBQ (potential and particularly those from the point of view of respondents) is a key characteristic that may impact respondents' cognitive processes and data quality. If the stimuli have similar meaning, the consequences for the cognitive process would not be very severe for the respondents' cognitive burden, as they might disregard one of the stimuli without a negative impact on measurement quality or provide a compound response without difficulties. However, as in the present data the stimuli mainly differed in the meaning, the author was not able to observe the potential impact of stimuli that are similar in meaning. More research on this topic is needed to support the above assumption. Nevertheless, one might also ask why researchers require the redundancy and unnecessary complexity associated with a DBQ with similar stimuli. If the meaning of stimuli is (or has a potential to be) different, the consequences might be more severe for both, respondents' cognitive process and data quality.

The remaining results could only be interpreted in the light of the former finding that the stimuli in DBQs are of rather different meaning. What are respondents doing in such a case? Do they disregard one of the stimuli or try to integrate both to a response? A frequent observation (for both inventories under investigation) was that the mean of one item with one stimuli of a DBQ did not differ from it notably, but there was a strong difference between the DBQ and the other stimulus. The author concludes from this observation that in a DBQ respondents attend to one of the stimuli (namely to that for which – if included in a SSQ – there was no mean difference from the DBQ). A less frequent observation was that respondents also tried to integrate the stimuli, as the mean of a DBQ differed from the other two versions (and the latter from each other). The strategy which respondents use and the reasons why they focus on one or other stimulus appears to be item specific (or rather content specific) and dependent on the salience of one word in a respondents' memory, on the familiarity of the word, or on its relevance to the concept from the point of view of respondents. More research is needed to understand these processes.

The next finding was surprising, because there were no differences in the time respondents took to respond to the DBQs or analogous questions with single stimuli, although there was often more text to read in a DBQ than in the questions with single stimuli. The hypothesis H2, that expected higher response times for DBQs due to the respondents' burden, was not corroborated. Some respondents even spent less time to respond to DBQs than to SSQs. Again, the results with respect to the response latencies were generalizable over both inventories. These results do not support previous findings by Bassili and Scott (1996), who observed higher response latencies for DBQs in their telephone study with students. As the present study was a self-administered online survey, the differences in findings could be due to mode difference. Next, the participants of the study by Bassili and Scott (1996) were highly educated, while the sample of the present study also contained less educated respondents. This is in line with the satisficing approach that expects satisficing behavior to be higher if the task difficulty is higher and respondents' cognitive abilities are lower (Krosnick 1991). Therefore, the presented findings with respect to the response times are consistent with the assumption of cognitive shortcuts by respondents who had to work on a difficult task. The results with respect to response times demonstrate respondents' focus on one of the stimuli in a DBQ, but not on both. Apart from the satisficing explanation that is in line with the observation of strongly comparable means between the DBQ and one of the SSQ versions (but not with the other), there are also alternative explanations available, that is, a stimulus is easier to understand in the context of other stimulus. However, this explanation is not plausible in the light of the results obtained for the mean differences and measurement invariance: Why is the version with one stimulus comparable to DBQ but not the other version? If respondents evaluate the double stimuli (with different meanings) in the context of each other, all three versions should differ in means and latent variable structures (a result found for one item but not for other items).

The results of the present study therefore support textbook arguments that the different parts of compound DBQs have a different meaning and that respondents consider one of the parts and disregard the other (e.g., Bradburn et al. 2004; Le Payne 1951; Oppenheim 1992).

With respect to the negative impact of the DBQs on data quality, no significant effect on reliability is observed, so that initially low reliabilities did not increase when using SSQs. This means that stimuli in DBQs (although different in meaning) did not have a potential to increase reliability, neither together, nor separated. Therefore, results may differ, if more reliable instruments with DBQs were revised.

However, the use of a DBQ or one of its parts, in fact, affected validity scores. The results show that convergent validity can increase when respondents have to evaluate one stimulus in each question and not two in DBQs. However, a strong and unexpected result was a gap in validity scores between the two single stimuli versions. This result shows that (if the stimuli are different in meaning) keeping one stimulus of an instrument that used to be a DBQ would be associated with a serious validity loss, which resulted for both inventories. Removing irrelevant stimuli is a frequent suggestion for repairing a DBQ (Olson 2008). The present results show that stimuli differ in meaning, also with respect to the pertinent latent construct, and one might not be aware of this difference or of whatever stimulus is a more optimal or prominent measure of the latent variable. Therefore, deleting a stimulus from a DBQ without knowing what stimulus in the DBQ is construct relevant would be a risky operation. For example, one can revise the items on the Schwartz's values assuming that more specific items were easier to respond to (according to the corresponding suggestions in textbooks, e.g., Dillman et al. 2014). As a consequence, one might use an instrument containing only the second sentences of the original PVQ (like the SSQ2 group in the present data). This researcher would then be very surprised if he or she obtains no significant correlations to the relevant third variables. There is also a risk of arriving at non-significant results for tests of hypotheses. The results of this study imply that one should be cautious when deleting stimuli in a DBQ. If a DBQ has to be revised, both stimuli in the pertinent two single questions should be tested in split ballot experiments or presented subsequently and examined with respect to possible improvement of validity. Taking into account the effort and costs of such revisions, it would be better to avoid the DBQs in questionnaire design from the outset.

An alternative possibility to revise a DBQ is to generate a single item for each of the stimuli and use them together in an inventory. While such a revision is easy to implement for PVQ (that means evaluation of each of the sentences as separate items), it is much more difficult to do so for other inventories such as B-RWA-6. Consider an inventory consisting of items: (1) “We should be grateful for leaders who tell us exactly what we shall do”; (2) “We should be grateful for leaders who tell us exactly how we shall do something.”; (3) “The age in which obedience to authority is one of the most important virtues should be over”; (4) The age in which discipline is one of the most important virtues should be over”. . . and so on. This means, having an item for each stimulus of a DBQ in one version requires significant rewording of the second sentence to avoid repeating contents in an item set. This calls for a strong change to the items what would have also introduced a higher error variance (in an experimental setting) and was the reason why the author of present article has not looked at the possibility of building a sentence for each stimulus in an inventory. Another disadvantage of such a revision for the B-RWA-6 (that is typical for inventories with DBQs in the social sciences and psychology) is that a twice as long questionnaire is needed, which also increases the complexity of respondents’ work and is less economic. Hence, it is possible that for the PVQ a revision in which each sentence is presented as a separate item is sensible. Further research should address this question.

Several limitations of the presented study require attention. Owing to the fact that a commercial assess panel was used, the generalizability of the presented results to other settings is restricted. The present research addressed the DBQs, in which the stimuli were connected with “and” or with comparable grammatical constructions. The results therefore do not apply to other complex questions, such as presuppositions or use of “or”. Two inventories were also involved, so that more research with other inventories and concepts, that is, behavioral questions, is needed. Although the two inventories differ in regards to the question form and complexity of DBQs, strongly comparable results emerged for both inventories. This allows similar results to be expected for other kinds of DBQs in which stimuli are enumerations or grammatically independent parts of a compound entity.

The results of the present study are in line with the assumption that respondents evaluate the double stimuli of a DBQ differently. The results also reflect that there would be a potential negative impact of DBQs on data quality as compared with one of the forms containing a single stimulus. Based on these results, the author shares the suggestion in the textbooks to avoid DBQs. In addition, the author suggests being cautious when revising existing DBQs. The author was surprised many times by the present results as the aim of the study was to empirically test a very old wisdom (a supposedly easy task). The straightforward expectations arising from this wisdom could only be supported with respect to the assumption that DBQs might include stimuli that are different in meaning. The expectation of a negative impact of DBQs and the positive one of the SSQs on response latencies and measurement quality was not as easy to show as initially expected. This seems to be because the stimuli in a DBQ would not only have a different meaning, but also different associations with the latent variable (as shown by the present research). The use of DBQs by respondents and their impact are less well understood and need more research, for example with respect to the negative impact of stimuli with similar meaning.

Further research should also address similarities and differences between DBQs and questions with presuppositions (Kay and Fillmore 1999) to shed more light on the cognitive processes behind them and to provide a more solid basis for advice on questionnaire design. Next, stimuli connected with "or" should be addressed as well, as the author is not aware of any empirical studies of this structure.

However, it is also important to consider whether it would be better to spend time and money to better understand the practice of using DBQs or if it would just be better to avoid them in inventories? Of course, the avoidance is difficult if established inventories are used, as was also clearly shown by the present article. However, if these inventories have sufficient reliability and validity, one can reuse them due to the lack of alternatives. If one tries to revise inventories (i.e., to adapt them to the current issues) or develop new inventories, the authors' strong advice is to avoid DBQs.

5. Appendix

5.1. Results for the Alignment Method

B-RWA-6

Notation:

DBQ: 1, SSQ1: 2; and SSQ2: 3.

Non-equivalent parameters are presented in brackets.

Loadings

AUT1 1 2 3
AUT2 1 (2) 3
AUT3 1 2 3
AUT4 1 2 3
AUT5 1 2 3
AUT6 1 2 3

Intercepts/Thresholds

AUT1 1 2 3
AUT2 1 (2) 3
AUT3 1 2 3
AUT4 1 2 3
AUT5 1 (2) 3
AUT6 (1) (2) (3)

Table A1. Factor mean comparison.

Class (group)	Latent mean	Groups with significantly smaller factor mean
SSQ2	0.182	–
DBQ	0.000	–
SSQ1	−0.006	–

5.2. Tradition and Conformity

Notation:
DBQ: 1, SSQ1: 2; and SSQ2: 3.
Non-equivalent parameters are presented in brackets.
Loadings
Trad1 1 (2) 3
Trad2 1 2 (3)
Conf1 1 2 3
Conf2 1 2 3
Intercepts/thresholds
Trad1 1 2 (3)
Trad2 1 (2) 3
Conf1 1 2 3
Conf2 1 2 3

Table A2. Factor mean comparison for tradition.

Class (group)	Mean	Groups with significantly smaller factor mean
SSQ2	0.000	SSQ1
DBQ	−0.229	SSQ1
SSQ1	−1.374	–

Table A3. Factor mean comparison for conformity.

Class (group)	Mean	Groups with significantly smaller factor mean
DBQ	0.327	SSQ1
SSQ2	0.000	–
SSQ1	−0.071	–

6. References

Aichholzer, J., and E. Zeglovits. 2015. “Balancierte Kurzskala autoritärer Einstellungen (B-RWA-6).” *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. DOI: https://doi.org/10.6102/zis239.

Altemeyer, B. 1981. *Right-Wing Authoritarianism*. Winnipeg: University of Manitoba Press.

Bassili, J.N., and B.S. Scott. 1996. “Response Latency as a Signal to Question Problems in Survey Research.” *Public Opinion Quarterly* 60(3): 390–399. DOI: https://org/doi.10.1086/297760.

Beauducel, A., and W.W. Wittmann. 2005. “Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure.” *Structural Equation*

Modeling: A Multidisciplinary Journal 12(1): 41–75. DOI: https://doi.org/10.1207/s15328007sem1201_3.

Beierlein, C., F. Asbrock, M. Kauff, and P. Schmidt. 2014. *Die Kurzskala Autoritarismus (KSA-3): Ein ökonomisches Messinstrument zur Erfassung dreier Subdimensionen autoritärer Einstellungen.* (GESIS-Working Papers, 2014/35). Mannheim: GESIS – Leibnitz-Institut für Sozialwissenschaften. Available at: http://www.gesis.org/fileadmin/kurzskalen/working_papers/KSA3_WorkingPapers_2014-35.pdf (accessed February 2019).

Bless, H., G. Bohner, T. Hild, and N. Schwarz. 1992. "Asking Difficult Questions: Task Complexity Increases the Impact of Response Alternatives." *European Journal of Social Psychology* 22(3): 309–312. DOI: https://doi.org/10.1002/ejsp. 2420220309.

Borgers, N., and J. Hox. 2001. "Item Nonresponse in Questionnaire Research with Children." *Journal of Official Statistics* 17(2): 321–335. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/item-nonresponse-in-questionnaire-research-with-children.pdf (accessed September 2020).

Bradburn, N.M., S. Sudman, and B. Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design: For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass.

Campbell, J.C., D.W. Webster, and N. Glass. 2009. "The Danger Assessment: Validation of a Lethality Risk Assessment Instrument for Intimate Partner Femicide." *Journal of Interpersonal Violence* 24(4): 653–674. DOI: https://doi.org/10.1177/08862605083doi.17180.

Chen, F.F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14(3): 464–504. DOI: https://doi.org/10.1080/10705510701301834.

Cohen, J. 1992. "A Power Primer." *Psychological Bulletin* 112(1): 155–159. DOI: https://doi.org/10.1037/0033-2909.112.1.155.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.

ESS, European Social Survey. 2014. *ESS Round 7 Source Questionnaire*. London: ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London. Available at: https://www.europeansocialsurvey.org/docs/round7/fieldwork/source/ESS7_source_main_questionnaire.pdf (accessed February 2019).

Fowler, F.J., Jr.. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56(2): 218–231. DOI: https://doi.org/10.1086/269312.

Gemenis, K. 2013. "Estimating Parties' Policy Positions through Voting Advice Applications. Some Methodological Considerations." *Acta Politica* 48(3): 268–295. DOI: https://doi.org/10.1057/ap. 2012.36.

Graesser, A.C. 2006. "Question Understanding Aid (QUAID): A Web Facility that Tests Question Comprehensibility." *Public Opinion Quarterly* 70(1): 3–22. DOI: https://doi.org/10.1093/poq/nfj012.

Grant Levy, S. 2019. "Deconstructing a Double-Barreled Alternative: Evolution and Creationism." *Psychological Reports* 122(5): 1995–2004. DOI: https://doi.org/10.1177/0033294118795145.

Groves, R.M., F.J. Fowler, Jr., M.P. Cooper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, (2nd ed.). Oxford: Wiley.

Hox, J.J., E.D. de Leeuw, and E.A.O. Zijlmans. 2015. "Measurement Equivalence in Mixed Mode Surveys." *Frontiers in Psychology* 6: 87. DOI: https://doi.org/10.3389/fpsyg.2015.00087.

Hu, L., and P.M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6(1): 1–55. DOI: https://doi.org/10.1080/10705519909540118.

Kane, M.T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50(1): 1–73. DOI: https://doi.org/10.1111/jedm.12000.

Kay, P., and C.J. Fillmore. 1999. "Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction." *Language* 75(1): 1–31. DOI: https://doi.org/10.2307/417472.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–216. DOI: https://doi.org/10.1002/acp. 2350050305.

Krosnick, J.A., and S. Presser. 2009. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by J.D. Wright and P.V. Marsden., (2nd ed.) (pp. 263–313). San Diego, CA: Elsevier.

Le Payne, S. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Lenzner, T., C. Neuert, P. Hadler, A. Stiegler, C. Beitz, R. Schmidt, Z. Umuc, N. Reisepatt, S. Andrea, and N. Menold. 2017. "Krankheitswissen und Informationsbedarfe – Diabetes mellitus. Fragebogen für Personen mit Diabetes." Available at: http://pretest.gesis.org/pdf/ProjektBericht/Projektbericht-17-03.pdf (accessed February 2019)

Lenzner, T., L. Kaczmirek, and A. Lenzner. 2010. "Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment." *Applied Cognitive Psychology* 24(7): 1003–1020. DOI: https://doi.org/10.1002/acp. 1602.

Lord, F.M., and M.R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.

Menold, N., and T. Raykov. 2015. "Can Reliability of Multiple Component Measuring Instruments Depend on Response Option Presentation Mode?" *Educational and Psychological Measurement* 76(3): 454–569. DOI: https://doi.org/10.1177/0013164415593602.

Meredith, W. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58(4): 525–543. DOI: https://doi.org/10.1007/BF02294825.

Messick, S. 1989. "Meaning and Values in Test Validation: The Science and Ethics of Assessment." *Educational Researcher* 18(2): 5–11. DOI: https://doi.org/10.3102/0013189X018002005.

Muthén, B.O. 2002. "Beyond SEM: General latent variable modeling." *Behaviormetrika* 29(1): 81–117. DOI: https://doi.org/10.2333/bhmk.29.81.

Muthén, B.O., and T. Asparouhov. 2014. "IRT Studies of Many Groups: The Alignment Method." *Frontiers in Psychology* 5,978. DOI: https://doi.org/10.3389/fpsyg.2014.00978.

Muthén, L.K., and B.O. Muthén. 2017. *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Olson, K. 2008. "Double Barreled Question." In *Encyclopedia of Survey Research Methods*, edited by P.J. Lavrakas. (pp. 209–211). Thousand Oaks, CA: Sage Publications. DOI: https://doi.org/10.4135/9781412963947.n145.

Oppenheim, A.N. 1992. *Questionnaire Design, Interviewing, and Attitude Measurement*. New York City: St. Martin's Press.

Raykov, T., and G.A. Marcoulides. 2011. *Introduction to Psychometric Theory*. New York: Taylor & Francis.

Raykov, T. 2012. "Scale Construction and Development Using Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by R.H. Hoyle. (pp. 472–492). New York: The Guilford Press.

Raykov, T., N. Menold, and G.A. Marcoulides. 2018. "Studying Latent Criterion Validity for Complex Structure Measuring Instruments Using Latent Variable Modeling." *Educational and Psychological Measurement* 78(5): 905–917. DOI: https://doi.org/10.1177/0013164417698017.

Reise, S. 2012. "The rediscovery of bifactor measurement models." *Multivariate Behavioral Research* 47: 667–696. DOI: https://doi.org/10.1080/00273171.2012.715555.

Schaeffer, N.C., and J. Dykema. 2011. "Questions for Surveys: Current Trends and Future Directions." *Public Opinion Quarterly* 75(5): 909–961. DOI: https://doi.org/10.1093/-poq/nfr048.

Schwartz, S.H., B. Breyer, and D. Danner. 2015. "Human Values Scale (ESS)." *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. DOI: https://doi.org/10.6102/zis234.

Schwartz, S.H. 2003. "A Proposal for Measuring Value Orientations across Nations." In *Questionnaire Development Package of the European Social Survey*. Available at: http://www.europeansocialsurvey.org/docs/methodology/core_ess_questionnaire/ESS_core_questionnaire_human_values.pdf (accessed February 2019).

Stafford, L. 2011. "Measuring relationship maintenance behaviors. Critique and development of the revised relationship maintenance behavior scale." *Journal of Social and Personal Relationships* 28(2): 278–303. DOI: https://doi.org/10.1177/0265407510378125.

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

Vettehen, P.G.H., and L.B. van Snippenburg. 2002. "Measuring Motivations for Media Exposure: A Thesis." *Quality and Quantity* 36(3): 259–276. DOI: https://doi.org/10.1023/A:1016076505379.

Williams, R.T., A.W. Heinemann, R.K. Bode, C.S. Wilson, J.R. Fann, and D.G. Tate. 2009. "Improving Measurement Properties of the Patient Health Questionnaire–9 with Rating Scale Analysis." *Rehabilitation Psychology* 54(2): 198–203. DOI: https://doi.org/10.1037/a0015529.

Yan, T., and R. Tourangeau. 2008. "Fast times and easy questions: the effects of age, experience and question complexity on web survey response times." *Applied Cognitive Psychology* 22(1): 51–68. DOI: https://doi.org/10.1002/acp. 1331.

Yorkston, K.M., C.R. Baylor, J. Dietz, B.J. Dudgeon, T. Eadie, R.M. Miller, and D. Amtmann. 2008. "Developing a Scale of Communicative Participation: A Cognitive Interviewing Study." *Disability and Rehabilitation* 30(6): 425–433. DOI: https://doi.org/10.1080/09638280701625328.

Received March 2019
Revised June 2020
Accepted August 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 887–906, http://dx.doi.org/10.2478/JOS-2020-0042

The Representativeness of Online Time Use Surveys. Effects of Individual Time Use Patterns and Survey Design on the Timing of Survey Dropout

Petrus te Braak[1], *Joeri Minnen*[1], *and Ignace Glorieux*[1]

Like other surveys, time use surveys are facing declining response rates. At the same time paper-and-pencil surveys are increasingly replaced by online surveys. Both the declining response rates and the shift to online research are expected to have an impact on the representativeness of survey data questioning whether they are still the most suitable instrument to obtain a reliable view on the organization of daily life. This contribution examines the representativeness of a self-administered online time use survey using Belgian data collected in 2013 and 2014. The design of the study was deliberately chosen to test the automated processes that replace interviewer support and its cost-efficiency. We use weighting coefficients, a life table and discrete-time survival analyses to better understand the timing and selectivity of dropout, with a focus on the effects of individual time use patterns and the survey design. The results show that there are three major hurdles that cause large groups of respondents to drop out. This dropout is selective, and this selectivity differs according to the dropout moment. The contribution aims to provide a better insight in dropout during the fieldwork and tries to contribute to the further improvement of survey methodology of online time use surveys.

Key words: Nonresponse; time use survey; survey design; online research.

1. Introduction

About 120 years ago, random sampling was proposed for population research (Bethlehem 2009). Before that time, elite consultations were often used for policy preparation. In particular socialist politicians perceived the survey sample as a democratic way to gauge the population for policy making (Savage and Burrows 2009). The aim of the random sample was to construct no less than a small society. Sample surveys were supposed to value everyone's opinion, regardless of one's socioeconomic status, and were therefore preferred over elitist policy making commissions.

Soon after the first practical applications of this approach, scholars started questioning nonresponse and the democratic nature of random sample surveys: with ever declining response rates (Curtin et al. 2000; Connelly et al. 2003; Cull et al. 2005; Johnson and Wislar 2012), that are disproportionally observed in groups of lower social background

[1] Vrije Universiteit Brussel, Department of Sociology, Research Group Tempus Omnia Revelat (TOR), Pleinlaan 2, 1050 Brussels, Belgium. Emails: ptebraak@vub.be, Joeri.Minnen@vub.be and Ignace.Glorieux@vub.be

Acknowledgments: We would like to thank the five anonymous referees for their careful and constructive reviews, which led to substantial improvements of this manuscript.

(Smith 2008; Porter and Whitcomb 2005), it is questionable whether sample surveys are still the democratic and representative method they once were supposed to be.

The same question applies to time use surveys. This approach has been historically developed to gain insight in the time use patterns for the purposes of planning of all sorts of organization of public life, such as working hours, traffic, household work, and sleeping time (Szalai 1966). This implies, of course, the inclusion of all social strata in society. Pääkkönen (1998) and Knulst and Van den Broek (1998), however, showed that time use surveys, too, are prone to selective nonresponse, which raises the question whether time use surveys are the right instrument to gain insight of the organization of public life. There may be an extra bias in such research, because a time use survey consists of several phases. Where "normal" survey research only consists of a questionnaire, a time use survey using a diary approach (of course this situation differs from other methodologies such as a telephone recall interview approach, used for the ATUS, see, for example Abraham et al. 2006) continues with a diary that needs to be kept over one or several days. The sequencing of these tasks, the atypical research method and the longer research period results in a higher participation burden (Pääkkönen 1998; Knulst and Van den Broek 1998). This leads to the question whether there may be two phases in which nonresponse takes place: (1) whether or not to accept the invite and complete the questionnaire; and (2) whether or not to complete the time diary after the questionnaire has been completed.

In addition to the differences between usual survey research and time use surveys, we have seen a second divide in methodology in recent years: the difference between paper-and-pencil surveys and online surveys. There is an interest to replace the traditional paper-and-pencil-(time use) survey by online surveys, in order to save costs, to speed up the fieldwork, and to avoid using interviewers who inevitably can cause bias. An emerging literature around nonresponse in online surveys shows that this method leads to extra bias due to its specific survey design (Couper et al. 2007; Smith 2008; Kwak and Radler 2002). The effects of this relatively new survey method on nonresponse, especially when used in a time use survey is, to our knowledge, still uncharted territory.

This article attempts to contribute to the debate on nonresponse in online time use surveys (OTUS) specifically. In 2013, an OTUS was conducted in Flanders (Belgium). In this contribution, we will elaborate on the timing of dropout during the fieldwork, and we will also discuss the selectivity of the dropout by (1) comparing the response during the questionnaire phase with other representative data from the same research population; and (2) by comparing the population of respondents of a successive survey with respondents of the base survey, to investigate the selectivity of dropout during the diary stage, as suggested by Porter and Whitcomb (2005) and Johnson and Wislar (2012).

The remainder of this article is structured as follows: firstly, we give an overview of existing literature on nonresponse in general and especially on nonresponse in time use surveys. Secondly, we describe the research design and data we use in order to shed light on the timing and selectivity of nonresponse in one of the first online time use surveys and formulate our hypotheses. Thirdly, we describe the results of the analyses. Lastly, we conclude by discussing the results in the light of the existing literature.

2. Literature Overview

Nonresponse bias in surveys can be related to the following aspects: background variables, behavior related to, or interest in, the research topic, or the survey design. In this section, these aspects are dealt with interchangeably, because there are often interrelationships between the aspects that lead to nonresponse. Studies on nonresponse bias in online surveys suggest that those research methods show in general the same nonresponse patterns as those of traditional paper-and-pencil surveys (Couper et al. 2007), although some other response patterns compared to paper-and-pencil surveys were found as well (Kwak and Radler 2002; Couper et al. 2007; Sax et al. 2003). For this literature review, we therefore concentrate mainly on variables related to time and on nonresponse in online surveys specifically.

In existing literature on nonresponse in time use surveys, often the focus is solely on the ultimate representativeness of such a survey, without taking into account the different phases of such a study. As far as we know, there is no existing literature about the possible extra bias caused by the additional nonresponse stemming from the diary phase in a time use survey. This makes it virtually impossible to know whether such a research design leads to an additional dropout and selectivity compared to a conventional research design. However, an extra bias would not be entirely illogical. Some groups might still be willing to fill in a questionnaire, but as soon as they experience the workload of a diary, they still drop out. Reasons for this may include, among others, the complexity, the degree of repetitiveness, but also time constraints. Our first hypothesis is therefore that *there are two different research phases in which nonresponse occurs: the first dropout occurs before finishing the questionnaire, the second when completing the diary* (H1). We expect that the nature of this dropout is selective and that the selectivity (partly) differs between the phases (i.e., that we see certain groups dropout more often during the questionnaire phase and for other covariates during the diary phase). We will formulate specific hypotheses about the dropout during the different stages for the different covariates.

In existing nonresponse literature, analyses generally show that socioeconomic status (SES) is one of the major causes of nonresponse bias in sample surveys of the general population (Porter and Whitcomb 2005). Others formulate this more specifically in terms of educational attainment: a lower level of education is related to lower levels of participation (Curtin et al. 2000; Singer et al. 1999). In the past, the higher nonresponse of the lower social strata was seen as a result of illiteracy (Goyder et al. 2002; Wallace 1954). Nowadays, illiteracy is not a problem with the same dimension anymore. However, as Goyder et al. (2002) make clear, mail surveys tend to have a higher SES bias compared to telephone surveys, indicating that probably the written language of paper-and-pencil surveys can still form a problem for some groups of potential respondents. One can argue that the same holds for OTUS, leading to the hypothesis that *the lower educated persons will have a lower response rate on the questionnaire of the OTUS than their higher educated counterparts* (H2q, for questionnaire), because the online context does not change the use of language in such a survey. However, language is less of a problem when filling in the diaries, because the structure of the questions is far less complex than for a questionnaire and the questions during the diary are very repetitive (what do you do, when, with whom?). In addition, a selection effect will have already taken place during the pre-questionnaire, as a result of which low-literate people may have already dropped out. As a result, we do expect *no additional dropout from the lower educated during the diary phase* (H2d, for diary).

In addition to SES, gender is usually identified as a significant predictor of survey participation (Curtin et al. 2000; Singer et al. 2000; Porter and Whitcomb 2005; Cull et al. 2005). Usually, scholars find that women are more willing to participate than men. However, there are also examples where no gender bias was found (Etter and Perneger 1997). Smith (2008) suggests that the willingness to participate is likely to be the result of how men and women make decisions. Female characteristics like empathy and emotional closeness are related, from this point of view, to survey participation. The overrepresentation of women is confirmed in some online surveys (Smith 2008; Sax et al. 2003). In another study, however, where Kwak and Radler (2002) compared mail and web-based surveys, they found that women were overrepresented in the mail survey and, surprisingly, underrepresented in the web-based survey (see also Dillman et al. 2009). The researchers explain this gender difference as men being more intensively involved with new technologies. Although this may have been the case in 2002, this hardly seems to be valid anymore. For these reasons, and in line with the discussed literature, *we hypothesize that women participate more in online time use surveys than men and that thus their dropout is lower during the questionnaire phase* (H3q)*, as well as during the diary phase* (H3d).

The relation between age and survey participation is, based on existing literature, rather unclear. Some find that younger age groups are more willing to participate (Goyder 1986; Moore and Tarnai 2001), whereas others find the opposite (Singer et al. 1999), or no relation between age and participation at all (Etter and Perneger 1997). In their overview of age effects, Groves and Couper (1998) indicate that there is slightly more support for the assumption that the refusal rate is higher for the elderly, but that effective nonresponse is not necessarily lower. In multivariate tests, the effect might be inverse. They hypothesize that the elderly have more civic duty, which would make them participate more often, at least for governmental surveys. Regarding online surveys, studies show that younger age groups are more willing to respond than their older counterparts (Kwak and Radler 2002; Couper et al. 2007), although Dillman et al. (2009) point out that the youngest and oldest age groups are underrepresented and that the middle groups are the ones that are overrepresented. The difference between age groups is sometimes explained on the basis of differences in computer ownership, frequency of internet use and IT literacy. Although for paper-and-pencil-surveys there seems to be little empirical or theoretical evidence to suggest a correlation between age and survey participation, these reasons provide sufficient arguments to assume that older age groups participate less often in online surveys. However, those who participate in the pre-questionnaire demonstrate through their participation that they have a computer, internet and the necessary skills to use them. In addition, we expect the elderly to show a little more civic duty. *We therefore expect older age groups to disproportionally drop out during the questionnaire phase* (H4q), *but once they have taken this step, there will be a selection effect, and that they therefore will drop out less than their younger counterparts during the diary phase of the fieldwork* (H4d) due to higher levels of civic duty.

Along with those typical background variables that are often associated with nonresponse in survey research, lack of interest in the research topic, or behavior related to the topic is often considered a cause for nonresponse. This is generally seen as a major threat for the validity of sample surveys. For time use surveys specifically, it is often feared that they suffer from specific nonresponse related to the use of time. In her monumental book on the time use of women, Hochschild and Machung (2003, 287) state that: "Ironically, the women most burdened by the very crunch the researchers were investigating, probably didn't have time to fill out such a lengthy questionnaire." According to Knulst and Van den Broek (1998), respondents need approximately 80 minutes to fill in a seven-day diary. If one adds the time one needs to respond to a pre- and post-questionnaire besides the self-administered time diary, it becomes clear that the average response burden in time use surveys using a diary approach is higher compared to usual survey research. In the existing literature on nonresponse in time use surveys, it is thus often assumed that particularly busy people do not have the time to participate in lengthy time use surveys (Groves and Couper 1998; Van Ingen et al. 2008; Pääkkönen 1998; Knulst and Van den Broek 1998; Abraham et al. 2006). Zuzanek (1998, 547) explicitly counters this reasoning by stating that: "busy people find time for all sorts of things, and are more likely to respond to time-diary questionnaires". A literature review of former studies finds empirical support for both perspectives. Whether or not they find an association between nonresponse and busyness seems to be dependent on the indicator used. Knulst and van den Broek (1998) measured busyness based on objective indicators as, for example, the time spent in work and other obligations, as do Abraham et al. (2006). Abraham et al. (2006) conclude that objective busyness does not have a major influence on response rates, based on the findings that full-time workers have higher response rates than part-time employees. They also find that people who work more than full-time have the same rates as people who work part-time (see also Pääkkönen 1998). In the Abraham et al. study, all groups have a higher response rate than people who do not work at all. This last finding is replicated by Knulst and van den Broek (1998). They find that the least busy groups are the most underrepresented, and conclude from this finding that the relation between being busy and response is somewhat surprising: busy people participate more often than people who are less busy.

Pääkkönen (1998) and Van Ingen et al. (2008) show that, using subjective indicators, busyness has no influence on the participation rate. In the research carried out by Pääkkönen (1998), respondents were asked whether or not they were in such a hurry that they did not manage to do everything they had to do and whether they had to give up things during regular weekdays because they did not have enough time before filling out the time diary. Both indicators of subjective busyness did not influence the willingness to participate in the time diary. Stress symptoms only had a minor negative effect on the participation rates of the diary in that study. The study carried out by Van Ingen et al. (2008) investigated the effect of the feeling of being in a rush, but did not find such an effect of this on participation either.

Based on these findings, we expect that busyness in objective terms is positively related to participation. We hypothesize that *busy people, such as the self-employed and full-time employees, have a lower dropout than their less busy counterparts, both during the questionnaire phase* (H5q) *and the diary phase* (H5d).

In addition to busyness, there are also other time-related factors that can have an effect on nonresponse. Nonresponse literature often focuses on the relationship between

Table 1. The hypotheses summarized

Timing	
Two different phases of dropout exist: 1) before filling in the questionnaire; 2) when completing the time diary	Hl
Education	
The low educated will have a lower response rate on the questionnaire than their higher educated counterparts	H2q
There will be no additional dropout from the lower educated during the diary phase	H2d
Gender	
Dropout of women is lower than the dropout of men during the questionnaire phase	H3q
Dropout of women is lower than the dropout of men during the diary phase	H3d
Age	
Older age groups will disproportionally drop out during the questionnaire phase	H4q
Older age groups will dropout less than their younger counterparts during the diary phase	H4d
Busyness	
Busy people have a lower dropout than their less busy counterpart during the questionnaire phase	H5q
Busy people have a lower dropout than their less busy counterpart during the diary phase	H5d
Deviating time patterns	
People wo deviate significantly in their use of time from the standard will be overrepresented in the questionnaire phase	H6q
People who deviate significantly in their use of time will be more inclined to continue to participate during the diary phase	H6d

response and interest in the research topic or topic salience (Marcus et al. 2007; Van Kenhove et al. 2000). Health surveys, for example, consistently report troublesome differences between participants and non-participants in health behavior (Boström et al. 1993; Smith and Nutbeam 1990; Hill et al. 1997) and the use of health services (Etter and Perneger 1997). Respondents with a specific profile and/or interest in the outcome of the study are often more willing to participate. This, of course, can cause problematic nonresponse bias when people who are interested have other attributes on the research topic than those with no interest. For time use surveys, this could mean that people who are aware that their time use differs from standard patterns, are more willing to respond. The reason would be that these people wish to be heard, that their deviant patterns are included in the figures. Therefore, we hypothesize that *people who deviate significantly in their use of time from the standard will be overrepresented in the questionnaire phase* (H6q) *and will be more inclined to continue to participate during the diary phase and therefore have a lower dropout* (H6d). All formulated hypotheses are summarized in Table 1.

3. Used Data Files

Before we go further into detail about our research design, we first describe the data and field work procedures in order to perform the nonresponse analysis. We make use of two different data sets.

3.1. TOR13

TOR13 is an OTUS that was conducted in 2013 and 2014 (Minnen et al. 2014). One of the objectives of TOR13 was to collect time use data by means of an innovative and cost-saving software platform that was developed inhouse. The goal was to replace paper-and-pencil methodology with a cost-reducing online methodology, and to test the functionality of the software platform. One of the greater savings was achieved by the introduction of automated fieldwork processes, and so the elimination of interviewers. In earlier time use surveys, the role of the interviewer was to convince respondents to participate, assist respondents with problems and register activities. In this study, this role was partly taken over by a web app that could be accessed via the browser. Convincing respondents to participate was done by letter. Invitations were sent out by post every two weeks in order to achieve a good distribution over a whole calendar year. If a respondent did not log in two weeks after sending out the invitation, a first reminder was sent. A second and last reminder was sent four weeks after the invitation. In case respondents had questions or needed help, this was only offered by email and telephone. The method made it possible to reduce the costs per respondent from EUR 260 to EUR 60–80. In total, a random sample of 39,756 people between 18 and 75 years living in Flanders (the Dutch-speaking part of Belgium) and Brussels (only Dutch-speakers were invited) was selected from the Belgian National Register with equal probabilities of being chosen. They were asked to complete a pre-questionnaire, keep a time diary for seven consecutive days and complete a post-questionnaire. In order to participate, one had to log in (using a username and password that were communicated via the letter) on a website where more information on the research project was offered and where they could start their participation. Those who did finish the research completely, automatically took part in a lottery where they could win different cash prizes with the maximum prize of EUR 500.

For 124 respondents, the gender and/or year of birth differed from the data in the national register. We consider this an indicator that a person other than the sampled person completed the questionnaire. For the analyses in this contribution, these people are excluded, since they were technically not sampled to participate in this study. Response rates will be discussed in detail in the results section of this contribution.

3.2. Labour Force Survey 2013

To evaluate the nonresponse bias on the TOR13 time use survey, we made use of the Belgian subsample of the Labour Force Survey (LFS). The LFS is an obligatory face-to-face survey of members of Belgian households that are at least 15 years old. Refusal is punished by law with a fine of EUR 40 to EUR 200. Despite its obligatory nature, this survey also suffers from some nonresponse due to inability to participate (for example, in the case of illness). Nevertheless, these data are generally known as the least bad option if one wishes to compare a sample with the total Belgian population and are therefore often used for this purpose. This study ran completely simultaneously with the TOR13 survey.

Despite the many similarities, some data operations are needed to make the data comparable. We restrict the research population of the LFS solely to the 18 to 75 years old inhabitants of Flanders (n = 37,828), in order to make the population comparable with that of the TOR13. The LFS does not allow making a distinction between Dutch- and

French-speaking Brusselians. Therefore, and because of the existing large differences in socioeconomic and cultural background between the Dutch- and French-speakers in Brussels, we decided to eliminate the Dutch-speaking inhabitants of Brussels from the TOR13 sample. After these data processings, the research populations can be considered as completely similar. The respondents analyzed in this study are thus residents of the Flemish Region between the ages of 18 and 75, all of whom were drawn individually on the basis of a random sample. A total of 36,665 sampled individuals (92.2%) of the TOR13 sample met these criteria.

4. Design

To check selective nonresponse and representativeness, we first use methods that come from demography and medical sciences. First, we use a life table (i.e., a mortality table)(Singer and Willett 2003, chap. 10) to determine at what point in the survey process respondents drop out. The sample population and the dropout rate during the survey process are very similar to a population and mortality in a life table. In the life table we will treat dropout as mortality. Different research phases are equated with reaching a certain age. More specifically, eleven research phases are taken into consideration: invitation, pre-questionnaire started, pre-questionnaire finished, logging at least one activity, logging one day, two days, three days, four days, five days, six days and finishing the time diary. Based on this table, we can then calculate the dropout probability in any particular phase in the research, as well as the cumulative survival.

Second, we will combine those methods with more regular approaches to analyse nonresponse. Porter and Whitcomb (2005) (see also Johnson and Wislar 2012), identify four such methods:

1. by comparing the response with other data sources,
2. by comparing panel respondents of a successive survey with respondents of a base survey,
3. by means of a nonresponse follow-up survey, and
4. by time of response analysis.

In this article, we will make use of the first two methods.

In order to map the selective nonresponse prior to the completion of the preliminary questionnaire, we will (by absence of own data at this point) compare the response on the pre-questionnaire with that of the Flemish subset of the Labour Force Survey 2013. Because of the many similarities between the two data sets, as mentioned earlier, this source offers the best available data for comparing the research population with the sampled population. We will use this method to analyze the nonresponse during the pre-questionnaire part of the survey. The two data sets will be compared using ratio weights, based on post stratification class adjustments. These are calculated as:

$$Ratio\,weights = \frac{realized\,sample\,(TOR13)\,in\%}{population\,(LFS13)\,in\,\%}$$

Weights higher than 1 indicate underrepresentation of the concerned subgroup in the TOR13 data, whereas weights below 1 indicate overrepresentation.

For the selective nonresponse in later stages of the fieldwork after finishing the pre-questionnaire, we will apply the method of comparing the base survey (i.e., the pre-questionnaire) to later stages (Porter and Whitcomb 2005; Johnson and Wislar 2012). For this, we fall back on the previously identified research phases. Because we can only rely on data from respondents who completed the questionnaire, eight research phases remain: finishing the pre-questionnaire (0), logging 1 activity (1), logging 1 day (2), 2 days (3), 3 days (4), 4 days (5), 5 days (6) and 6 days (7). By filling in the pre-questionnaire, respondents provided us with information on their background. For the nonresponse occurring during the time diary, we can thus execute a comparison between responders and non-responders on all those who finished the pre-questionnaire. We will do this using Discrete-Time Survival Analysis (Singer and Willet 2003, chap. 11 and 12). These analyses focus on describing whether and when events occur and whether or not differences can be observed in occurrence and timing to different covariates. For this, we will use standard logistic regression on a person-period data set. In a typical person data set, every person has one record or line of data, whereas in a person-period data set every person has multiple records. The exact amount of records is dependent on the periods that this person is at risk. For example, if someone drops out of the fieldwork in the third stage, the data set will have three records. In this case, the maximum amount of records is eight, since we take eight stages of the fieldwork into account. The indicator for time that will be used here is thus measured discretely. The event, or in other words the dependent variable, that is under investigation here is dropout, or, put differently, failing to finish the research.

5. Covariates That Will Be Taken Into Account

In order to analyze a potential selective nonresponse to the pre-questionnaire, we had to limit the categories per variable to the level of detail of the least detailed questionnaire. We use the following operationalization for the background variables: Gender (two categories: Male/Female), Age (five categories: 18 to 24 years/25 to 39 years/40 to 54 years/55 to 64 years/65 + years), and Education (three categories: ISCED 1 to 2 (low)/ISCED 3 to 4 (average)/ISCED 5 to 8 (high)).

In order to measure objective busyness, we rely on pre-questionnaire data concerning occupational status (six categories: working/unemployed/student/incapacitated/retired/other) and full versus part-time employment (four categories: self-employed/full-time/part-time/no contract) (i.e., Knulst and van den Broek 1998; Abraham, et al. 2006).

A deviating use of time will be measured using work schedule (five categories: Non-working, Fixed shifts, Flexible shifts, and other), Evening and Night shifts (three categories: Never/occasionally, Regularly, Always/non-working). All of these variables were answered by most of the respondents, except the questions on evening and night shifts. They were part of an extra sub-questionnaire that only had to be filled in by a randomly selected subsample of the total population (this was done in order to reduce participation burden) and are thus only answered by a *random* 40% of the total sample population.

6. Results

The results will be discussed in the following order: first, we will present the exact timing and occurrence of dropout during the fieldwork. Second, we will present the comparison

Table 2. Life table of the dropout during the TOR13-fieldwork

Status achieved at the start of the phase	Number Entering Interval	Censored	Dropout (in n)	Dropout (in %)	Survival (in %)	Cumulative survival (in %)
Invite	36,665	0	24,768	67.6	34.4	34.4
Pre-questionnaire started	11,897	0	776	6.5	93.5	30.3
Pre-questionnaire finished	11,121	0	4,427	39.8	60.2	18.3
1 activity logged	6,694	0	2,106	31.5	68.5	12.5
1 day logged	4,588	0	445	9.7	90.3	11.3
2 days logged	4,143	0	204	4.9	95.1	10.7
3 days logged	3,939	0	150	3.8	96.2	10.3
4 days logged	3,789	0	121	3.2	96.8	10.0
5 days logged	3,668	0	98	2.7	97.3	9.7
6 days logged	3,570	3,428	142	4.0	96.0	9.3

between responders on the pre-questionnaire and the LFS13. Last, the analyses of dropout of all those who finished the pre-questionnaire in later stages of the fieldwork will be shown using discrete-time survival analyses.

6.1. The Occurrence and Timing of Nonresponse

Table 2 shows the dropout during the different phases of the TOR13 fieldwork. There are a number of times when the dropout is high. Nearly seven in ten of the initial sample drop out before the questionnaire is completed, because they do not respond to the invite, or do not finish the pre-questionnaire. During the diary phase, another part of the original sample drops out. The data thus show support for H1. That being said, the actual situation is more complex than described in H1. The majority of the original sample fall out before the questionnaire has started, and the diary phase consists of several phases that lead to dropout. The most important dropout during the diary phase occurs when the first activity and the first day are registered, only then does it gradually decrease. In the next section, we will investigate to what extent this dropout is selective. We do this not only by making a distinction between the questionnaire phase and the diary phase. We will also consider the different phases of the diary stage when checking for selectivity in dropout. However, we will first discuss the selectivity of the 69.7% who drop out before finishing the questionnaire.

6.2. Nonresponse in the Pre-Questionnaire Phase

The selectivity of the dropout prior to the diary phase is analyzed by using a comparison with the mandatory Labour Force Survey 2013 (see Table 3. The ratios in the TOR13 data are calculated on the entire population living in Flanders who completed the preliminary questionnaire (n = 11,121), while the ratios in the LFS 2013 are calculated on all 18 to 75-year-old residents of Flanders who participated (n = 37,828) A weighting coefficient higher than 1 indicates an underrepresentation in the TOR13 data, whereas a coefficient lower than 1 indicates an overrepresentation of the subpopulation in question.

Table 3. Composition of the population. the realized sample for the pre-questionnaire and the ratio weights

		Population in %* (LFS13) [CI 95%]	Realized sample in %* (TOR13) [CI 95%]	Ratio weights
Gender				
	Female	50.5 [50.0-51.0]	53.0 [52.1-53.9]	0.95
	Male	49.5 [49.0-50.0]	47.0 [46.1-47.9]	1.05
Age				
	18 to 24 years	11.1 [10.8-11.4]	13.4 [12.8-14.0]	0.83
	25 to 39 years	22.6 [22.2-23.0]	26.1 [25.3-26.9]	0.87
	40 to 54 years	30.4 [29.9-30.9]	31.3 [30.4-32.2]	0.97
	55 to 64 years	18.8 [18.4-19.2]	18.7 [18.0-19.4]	1.01
	65+ years	17.0 [16.6-17.4]	10.5 [9.9-11.1]	1.62
Education				
	Low (ISCED 1 to 2)	29.7 [29.2-30.2]	22.2 [21.4-23.0]	1.34
	Average (ISCED 3 to 4)	39.7 [39.2-40.2]	35.9 [35.0-36.8]	1.11
	High (ISCED 5 to 8)	30.6 [30.1-31.1]	41.9 [41.0-42.8]	0.73
Occupational status				
	Working	56.4 [55.9-56.9]	64.7 [63.8-65.6]	0.87
	Unemployed	4.6 [4.4-4.8]	3.6 [3.3-3.9]	1.28
	Student	6.3 [6.1-6.5]	9.1 [8.6-9.6]	0.69
	Incapacitated	4.9 [4.7-5.1]	2.1 [1.8-2.4]	2.33
	Retired	22.2 [21.8-22.6]	17.0 [16.3-17.7]	1.31
	Other	5.6 [5.4-5.8]	3.5 [3.2-3.8]	1.60
Full/part-time employment				
	Non-working	43.4 [42.9-43.9]	35.5 [34.6-36.4]	1.22
	Self-employed	9.0 [8.7-9.3]	7.5 [7.0-8.0]	1.20
	Full-time	34.3 [33.8-34.8]	42.6 [41.743.5]	0.81
	Part-time	13.3 [13.0-13.6]	14.4 [13.7-15.1]	0.92
Work schedule				
	Non-working	44.4 [43.9-44.9]	36.4 [35.5-37.3]	1.22
	Fixed schedule	39.3 [38.8-39.8]	24.4 [23.6-25.2]	1.61
	Flexible schedule	13.5 [13.2-13.8]	36.0 [35.1-36.9]	0.38
	Other schedule	2.8 [2.6-3.0]	3.3 [3.0-3.6]	0.85
Evening shifts				
	Non-working	44.4 [43.9-44.9]	36.5 [34.9-38.1]	1.22
	Never	35.9 [35.4-36.4]	24.6 [23.2-26.0]	1.46
	Occasionally/ regularly/always	19.7 [19.3-20.1]	38.9 [37.3-40.5]	0.51
Night shifts				
	Non-working	44.4 [43.9-44.9]	36.5 [34.9-38.1]	1.22
	Never	49.1 [48.6-49.6]	45.9 [44.3-47.5]	1.07
	Occasionally/ regularly/always	6.6 [6.3-6.9]	17.6 [16.4-18.8]	0.38

*Column percentages per individual variable sum up to 100%

The different age groups show some selective nonresponse bias. In general, we find that the lower the age, the better the chance of participation. Up to the age of 54, we see an overrepresentation in the data, while groups from the age of 55 are underrepresented in the sample population. From the age of 65 or older, the underrepresentation becomes even larger. This is possibly a result of the online setup of the survey instrument. This age group is known for having less ICT competences and a lower accessibility to internet connections. All other age groups are somewhat overrepresented in the sample. Therefore, we consider H4q not rejected.

Rather large differences in participation in the pre-questionnaire are also found between the educational classes. The results show that the higher the level of education, the higher the final participation will be. The higher educated are strongly overrepresented, whereas the lower educated are underrepresented. People with an average level of education are only slightly underrepresented. This finding is in accordance with H2q.

Based on occupational status, we find that, consistent with findings on nonresponse bias in usual survey research, the incapacitated are strongly underrepresented. Also, retirees and unemployed persons are a little underrepresented. Furthermore, we find that students, and to a smaller extent people who are working, are rather overrepresented. The full/part-time employment divide leads to minor differences in response. Both people working on a full-time basis and on a part-time basis are overrepresented. This is not surprising because both form a large part of the group that are currently employed, that we have previously seen as being overrepresented. Although the differences are small, we see that people who work part-time are slightly less likely to participate in the OTUS than people who work full-time. The self-employed, a category that cannot be put in either the full or part-time categories, are slightly underrepresented. We consider these findings as more complex than H5q suggests, namely that normally busy people do indeed participate slightly more often than people with a lower or no work-related time pressure, but that this does not apply to the very busiest, in this contribution the self-employed.

The response differs based on the kind of work schedule of the respondent. The largest share of the population, people with a fixed working schedule, are somewhat underrepresented, whereas people with flexible hours (working in shifts, those who have a flexible schedule or an intermittent schedule) are overrepresented in our survey sample compared to the total Belgian population. The same holds for people who work evening and night shifts. These groups are strongly overrepresented in the sample. These findings thus show full support for H6q, namely that people with deviating, non-standard working hours do indeed seem to participate more often in a questionnaire about time use.

Lastly, small differences in participation on the pre-questionnaire between men and women were found. Hence, we consider H3q as not refuted.

6.3. *Nonresponse in Later Stages of the Survey*

After filling in the pre-questionnaire, respondents should start with keeping the seven-day diary. In this part, we can rely on data that were reported by the respondents themselves during the pre-questionnaire. Models 1a and 1b in Table 4 show the analyses on the complete subsample of those who filled in the pre-questionnaire. In order to cope with the problem that only a random 40% of the sample filled in the questions on evening and night shifts, we show

Table 4. Discrete-Time Survival Analysis on dropout in the time diary keeping phase

	Model 1a				Model 1b				Model 2a				Model 2b			
			[CI 95%]				[CI 95%]				[CI 95%]				[CI 95%]	
	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper
Constant	0.636	***			0.585	***			0.598				0.569	***		
Time (ref. pre-question naire finished)																
1 activity	0.697	***	0.652	0.745	0.716	***	0.670	0.767	0.659	***	0.587	0.740	0.683	***	0.608	0.768
1 day	0.164	***	0.147	0.183	0.171	***	0.153	0.191	0.163	***	0.135	0.197	0.171	***	0.142	0.207
2 days	0.084	***	0.072	0.097	0.088	***	0.076	0.102	0.098	***	0.077	0.124	0.103	***	0.081	0.131
3 days	0.062	***	0.052	0.074	0.065	***	0.055	0.077	0.067	***	0.050	0.089	0.071	***	0.053	0.094
4 days	0.051	***	0.043	0.062	0.054	***	0.044	0.065	0.069	***	O.051	0.091	0.072	***	0.054	0.097
5 days	0.043	***	0.034	0.053	0.045	***	0.036	0.055	0.053	***	0.038	0.073	0.056	***	0.040	0.077
6 days	0.063	***	0.052	0.075	0.066	***	0.055	0.079	0.062	***	0.045	0.084	0.066	***	0.048	0.090
Gender (ref.: female)																
Male					1.122	***	1.055	1.193					1.120	*	1.007	1.245
Age (40 to 54 years)																
18 to 24 years					1.232	**	1.079	1.407					1.293	*	1.034	1.617
25 to 39 years					1.126	**	1.043	1.216					1.123	n.s.	0.983	1.282
55 to 64 years					0.982	n.s.	0.890	1.082					1.026	n.s.	0.871	1.209
65 + years					1.345	***	1.144	1.582					1.526	**	1.156	2.015
Education (ref.: ISCED 3 to 4)																
ISCED 1 to 2					1.401	***	1.293	1.517					1.452	***	1.267	1.664
ISCED 5 to 8					0.734	***	0.686	0.785					0.734	***	0.654	0.825
Occupational status (ref.: working)																
Unemployed					1.112	n.s.	0.943	1.311					1.206	n.s.	0.900	1.616
Student					0.715	***	0.613	0.834					0.644	**	0.487	0.852
Incapa citated					1.231	n.s.	0.998	1.518					1.049	n.s.	0.737	1.491
Retired					0.832	*	0.722	0.959					0.708	**	0.546	0.920
Other					1.047	n.s.	0.885	1.238					0.962	n.s.	0.711	1.302

Table 4. Continued

	Model 1a				Model 1b				Model 2a				Model 2b			
			[CI 95%]				[CI 95%]				[CI 95%]				[CI 95%]	
	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper	OR	Sig.	Lower	Upper
Full/part-time employ ment (ref.: full-time employment)																
Self-employed					1.423	***	1.266	1.599					1.555	***	1.270	1.903
Part-time					1.078	n.s.	0.982	1.184					0.997	n.s.	0.848	1.173
Work schedule (ref.: fixed schedule)																
Flexible schedule					1.029	n.s.	0.953	1.112					1.030	n.s.	0.897	1.183
Evening shift (ref.: never)																
Occasionally/regularly/always													0.927	n.s.	0.800	1.074
Night shift (ref.: never)																
Occasionally/regu larly/always													1.028	n.s.	0.881	1.198
-2LL	29,678.166				29,297.752				10,337.141				10,175.268			
df	7				22				7				24			
Sig.						***				***						
Nagelkerke R2	.262				.276				.243				.260			
N episodes	38,687				38,687				13,744				13,744			

the analyses of those variables separately in models 2a and 2b. The time indicators bear the name of the initial status. For example, the “1 day” indicator shows the odds ratio (OR) on which someone drops out in the interval between one and two days logged.

In this part, due to a lack of space, we only show the analyses about the effect of the different covariates on dropout during the entire diary phase to test our hypotheses. As was previously made clear, however, the relation between dropout and the covariates during the diary phase can differ from phase to phase. The models that take this into consideration, in which interactions between the time indicators and the covariates are included, would take up too much space and can thus be found in Supplemental data. Where they provide useful insights, they will be discussed below.

Models 1a and 2a show the plain effect of time on nonresponse. We do not go into detail on those effects, because they are already discussed in a clear way using the life table analysis. Model 1b shows a better fit than its predecessor, indicating that the covariates offer additional explanation power for the chances of a dropout. This can be seen as an indicator that selective nonresponse also occurs in the diary phase.

Men are 1.122 times more likely than women to drop out during the diary phase. These results can be considered as supportive for H3d. The interaction terms (see Supplemental data online, Table 1) show that this difference mainly occurs in the starting phases of the diary. The differences are largest before one activity was registered (OR = 1,266). The difference then gradually decreases. After three days have been registered, there is no longer a difference in dropout between the genders.

Regarding age, the results show that both the youngest groups (up to 39 years) and the oldest group (65+) have a higher dropout during the diary phase than the 40 to 64-year-olds. The dropout is highest among the over-65s (OR = 1.345), followed by the 18 to 24-year-olds (OR = 1.232) and the 25 to 39-year-olds (OR = 1.126). H4d can thus neither be accepted nor rejected. The middle age groups clearly drop out the least. The interaction effects, however, between time and age show interesting differences. The oldest age group who already dropped out excessively during the questionnaire, and who it was suspected would drop out less often during the diary phase, experience a high dropout during the first phases of the diary. The previously described selection effect only takes place after these respondents have demonstrated that they can handle the diaries by filling in at least one activity. This is different for younger age groups (up to 39 years old). They take the first hurdle more easily (registering the first activity) and then drop out more often during the first, second, third and fourth day of the fieldwork (see Supplemental data online, Table 2). It remains unclear why they dropout later. It is possible that this group will drop out due to the repetitive nature of a time use survey.

The models also show fairly large differences in dropout between the low and highly education. The lower educated dropout more often than the middle educated (OR = 1.401), while highly educated have a lower odds ratio (OR = 0.734). H2d thus has to be clearly refuted. The interaction effects show that, here too (Supplemental data online, Table 3), the effects are greatest at the start of the diary phase. It appears that primarily people with a low level of education drop out before completing the first activity and the first day. During the later days, this difference between the educational groups will gradually decrease.

Regarding professional status, we see no major differences in dropout between the working, the unemployed, the incapacitated and others. We do, however, observe differences with students (OR = 0.715) and the retired (OR = 0.832). Both groups have a lower dropout during the diary phase than the others. The interaction effects for the retired show the same pattern as those for the oldest age group (see Supplemental data online, Table 4). They dropout before completing the first activity, after which the differences become smaller. Ultimately, they will thus drop out less than the other groups. The incapacitated persons also experienced problems with the registration of the first activity. Unlike the retired persons, the participation of this group does not seem to improve significantly after they have taken this first hurdle. As a result, their final dropout will be slightly higher than that of the other occupational groups (see Supplemental data online, Table 4, model 2). When controlled for other variables, this difference vanishes. The lower dropout rate among students mainly occurs at the start of the diary phase and they come more often than others through the phase in which the first activity must be registered. Only during the fourth day of registration is their dropout slightly higher than for other groups. Based on the number of hours worked, we see no difference between part-time and full-time workers. It is clear, however, that the self-employed drop out more often than people in paid employment. This difference only occurs during the registration of the first activity, in which the self-employed drop out more often. Based on these findings, it is impossible to assume that busy groups will drop out less during the diary phase. Therefore, we also reject H5d. Non-working people do not differ so much from working people in terms of dropout. No differences were found either between full-time or part-time workers. However, it is found that retirees (even checked for age) drop out less often and that the self-employed drop out more often. Specifically for the diary phase, these findings indicate that people with a busy work schedule have a slightly higher dropout rate than people with a quieter work schedule. This finding, however, should be further investigated.

No differences were found in dropout based on work schedule and whether or not evening and night shifts work (see model 2b). These results can be considered as supportive for H6d, which states that people with a deviant work schedule do not drop out more or less frequently than people with a more regular work schedule.

7. Discussion

In this contribution, we examined both the timing of dropping out, and the selectivity thereof in online time use surveys using a diary approach. By making use of methods such as life tables and discrete-time survival analysis, where dropout is considered the transition that someone (preferably not) encounters during the duration of a complete time use survey using diaries, we found that timing of dropout is related to selectivity of this dropout. Different groups show different behavior during the course of the study. Their dropout varies, not only in the final result, but also in the exact timing.

The results showed that there are two (or, depending on the definition, three) major moments of dropout during the course of a typical online diary research. The largest dropout occurs before one question has been answered. In that respect, time use research is no different than normal surveys. Almost 68% of the invited respondents dropped out before starting the questionnaire. What may be different in this study, however, is the reason for dropping out. These dropouts include not only people who do not wish to participate, but also people who do not have internet access or the necessary IT competences to participate in an online study. Two other major hurdles where many respondents dropped out were the phases in which respondents had to fill in the first activity and the first day, both at the very start of the diary phase. After that, dropout decreases steadily.

Moreover, the phases in which dropout mainly occurs show a selective dropout. The selectivity on the questionnaire leads to a first underrepresentation of over-65s and the low and middle-educated, very similar to the results that were found by Dillman et al. (2009). In addition, we found that men, the self-employed, non-working persons, such as unemployed persons, incapacitated persons and retired persons are underrepresented in this phase. Also noteworthy is the overrepresentation of people who have a flexible work schedule and of those who work evening and night shifts.

Where normal survey research generally ends after the questionnaire, the second and arguably the most important part of the time use survey begins. This second phase, especially the start of it, leads to an additional dropout of the subsample that finished the questionnaire. This dropout is also selective. To a large extent, this selectivity is in the same direction as with regard to the selectivity on the questionnaire. An additional group of men drop out extra often, just like the over-65s, the lower educated and the self-employed. Groups such as the unemployed and the incapacitated participate less in the questionnaire, but once they do participate, they do not show additional dropout.

With regard to the interaction between timing and selectivity, it is noticeable that a high degree of selective dropout occurs especially during the first two days of the diary keeping. For example, it turns out that mainly over-65s, the retired, the incapacitated and the self-employed often fall out disproportionately while completing the first activity and the first two days. A seven-day diary therefore does not lead to extra selectivity compared to a two-day diary.

The specific nonresponse at the start of the diary phase should thus deserve attention in future time use surveys using a diary approach. The elderly, the retired and the incapacitated may be groups that need additional help with the first steps in which the diary is completed. In other surveys, this is where the interviewer steps in. Possibly, an interviewer is able to lower the dropout of those subpopulations. This may also apply to the self-employed, but it seems more likely that they will drop out as soon as they experience the high level of participation burden. To counteract this selectivity in dropout, consideration can be given to motivating this group extra by responding to their specific use of time, as well as offering them (higher) compensation for their time loss. However, we expect that it would be more effective, not only for this group, but for all participants, to structurally reduce the participation burden typical of time use surveys.

For this, we expect a lot from the shift from web surveys to surveys on mobile devices. Before the introduction of the web survey, respondents were given a paper diary that they could fill in at any time of the day. This changed at the introduction of the web survey where people had to have a computer (and internet connection) at their disposal, which replaced direct registration with retrospective registration (usually during the evening), and, in turn, increased the response burden, because of the effort to reconstruct the day. The introduction of registration via mobile devices again makes it possible to actively fill in a diary at any time. Besides, it offers possibilities for imputing diaries with passive data (from smartphones, watches, etc.), leading to further reduction of the participation burden. This could also help to counter the slightly higher dropout rates of groups with high IT competences (such as students in this study) during later phases of the fieldwork. They may drop out due to the higher degree of repetition of the diaries. Own recent, unpublished, research shows that people nowadays often use a combination of devices for their participation. Future research will have to clarify how this shift affects representativeness.

Furthermore, future research about nonresponse in surveys should make clear to what extent the selective dropout leads to bias on time use parameters. Hence it is necessary, certainly for time use surveys, to create clarity about the relationship between busyness and dropout. Another research track that should be followed is how dropout can be prevented at the different identified hurdles by improving the research methodology.

8. References

Abraham, K.G., A. Maitland, and S.M. Bianchi. 2006. "Nonresponse in the American Time Use Survey: Who Is Missing from the Data and How Much Does It Matter?" *Public Opinion Quarterly* 70(5): 676–703. DOI: https://doi.org/10.1093/poq/nfl037.

Bethlehem, J. 2009. "The Rise of Survey Sampling." *Statistics Netherlands*. Discussion Paper (09015). Available at: https://www.cbs.nl/-/media/imported/documents/2009/07/2009-15-x10-pub.pdf?la = nl-nl&hash = B75A64DF0877B7FD089E796FCFE81145 (accessed August 2020).

Boström, G., J. Hallqvist, B.J.A. Haglund, A. Romelsjö, L. Svanström, and F. Diderichsen. 1993. "Socioeconomic Differences in Smoking in an Urban Swedish Population. The Bias Introduced by Non-Participation in a Mailed Questionnaire." *Scandinavian Journal of Public Health* 21(2): 77–82. DOI: https://doi.org/10.1177/140349489302100204.

Connelly, N.A., T.L. Brown, and D.J. Decker. 2003. "Factors Affecting Response Rates to Natural Resource – Focused Mail Surveys: Empirical Evidence of Declining Rates Over Time." *Society & Natural Resources* 16(6): 541–549. DOI: https://doi.org/10.1080/08941920309152.

Couper, M.P., A. Kapteyn, M. Schonlau, and J. Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36(1): 131–148. DOI: https://doi.org/10.1016/j.ssresearch.2005.10.002.

Cull, W.L., K.G.O. Connor, S. Sharp, and S.S. Tang. 2005. "Methods Response Rates and Response Bias for 50 Surveys of Pediatricians." *HSR: Health Services Research* 40(1): 213–226. DOI: https://doi.org/10.1111/j.1475-6773.2005.00350.x.

Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64(4): 413–428. DOI: https://doi.org/10.1086/318638.

Dillman, D.A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B.L. Messer. 2009. "Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet." *Social Science Research*, 38(1), 1–18. DOI: https://doi.org/10.1016/j.ssresearch.2008.03.007.

Etter, J.-F., and T.V. Perneger. 1997. "Analysis of Non-Response Bias in a Mailed Health Survey." *Journal of Clinical Epidemiology* 50(10): 1123–1128. DOI: https://doi.org/10.1016/S0895-4356(97)00166-2.

Goyder, J. 1986. "Surveys on Surveys: Limitations and Potentialities." *Public Opinion Quarterly* 50(1): 27–41. DOI: https://doi.org/10.1086/268957.

Goyder, J., K. Warriner, and S. Miller. 2002. "Evaluating Socio-Economic Status (SES) Bias in Survey Nonresponse." *Journal of Official Statistics* 18(1): 1–11. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/evaluating-socio-economic-status-ses-bias-in-survey-nonresponse.pdf (accessed September 2020).

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York (NY): John Wiley. DOI: https://doi.org/10.1525/aa.1999.101.3.699.

Hill, A., J. Roberts, P. Ewings, and D. Gunnell. 1997. "Non-Response Bias in a Lifestyle Survey." *Journal of Public Health* 19(2): 203–207. DOI: https://doi.org/10.1093/oxfordjournals.pubmed.a024610.

Hochschild, A., and A. Machung. 2003. *The Second Shift: Working Families and the Revolution at Home*. New York (NY): Penguin.

Johnson, T.P., and J.S. Wislar. 2012. "Response Rates and Nonresponse Errors in Surveys." *JAMA: The Journal of the American Medical Association* 307(17): 1805–1806. DOI: https://doi.org/10.1001/jama.2012.3532.

Knulst, W., and A. van den Broek. 1998. "Do Time-Use Surveys Succeed in Measuring 'Busyness'? Some Observations of the Dutch Case." *Loisir et Société / Society and Leisure* 21(2): 563–572. DOI: https://doi.org/10.1080/07053436.1998.10753671.

Kwak, N., and B. Radler. 2002. "A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality." *Journal of Official Statistics* 18(2): 257–273. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-comparison-between-mail-and-web-surveys-response-pattern-respondent-profile-and-data-quality.pdf (accessed September 2020).

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schutz. 2007. "Compensating for Low Topic Interest and Long Surveys: A Field Experiment on Nonresponse in Web Surveys." *Social Science Computer Review* 25(3): 372–383. DOI: https://doi.org/10.1177/0894439307297606.

Minnen, J., I. Glorieux, T.P. van Tienoven, D. Weenas, J. Deyaert, S. van den Bogaert, and S. Rymenants. 2014. "Modular Online Time Use Survey (MOTUS) – Translating an existing method in the 21st century." *Electronic International Journal of Time Use Research*, 11(1), 73–93. DOI: https://doi.org/10.13085/eIJTUR.11.1.73-93.

Moore, D.L., and J. Tarnai. 2001. "Evaluating Nonresponse Error in Mail Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 197–212. New York (NY): Wiley.

Pääkkönen, H. 1998. "Are Busy People Under- or over-Represented in National Time Budget Surveys?" *Loisir et Société / Society and Leisure* 21(2): 573–582. DOI: https://doi.org/10.1080/07053436.1998.10753672.

Porter, S.R., and M.E. Whitcomb. 2005. "Non-Response in Student Surveys: The Role of Demographics, Engagement and Personality." *Research in Higher Education* 46(2): 127–152. DOI: https://doi.org/10.1007/s11162-004-1597-2.

Savage, M., and R. Burrows. 2009. "Some Further Reflections on the Coming Crisis of Empirical Sociology." *Sociology* 43(4): 762–772. DOI: https://doi.org/10.1177/0038038509105420.

Sax, L.J., S.K. Gilmartin, and A.N. Bryant. 2003. "Assessing Response Rate and Nonreponse Bias in Web and Paper Surveys." *Research in Higher Education* 44(4): 409–432. DOI: https://doi.org/10.1023/A:1024232915870.

Singer, E., R.M. Groves, and A.D. Corning. 1999. "Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation." *Public Opinion Quarterly* 63(2): 251–260. DOI: https://doi.org/10.1086/297714.

Singer, E., J. van Hoewyk, and M.P. Maher. 2000. "Experiments with Incentives in Telephone Surveys." *Public Opinion Quarterly* 64(2): 171–188. DOI: https://doi.org/10.1086/317761.

Singer, J.D., and J.B. Willett. 2003. *Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence*. New York (NY): Oxford University Press.

Smith, C., and D. Nutbeam. 1990. "Assessing Non-Response Bias: A Case Study from the 1985 Welsh Heart Health Survey." *Health Education Research* 5(3): 381–386. DOI: https://doi.org/10.1093/her/5.3.381.

Smith, W.G. 2008. "Does Gender Influence Online Survey Participation? A Record-Linkage Analysis of University Faculty Online Survey Response Behavior." Available at: https://files.eric.ed.gov/fulltext/ED501717.pdf (accessed August 2020).

Szalai, A. 1966. "Trends in Comparative Time-Budget Research." *American Behavioral Scientist* 9(9): 3–8. DOI: https://doi.org/10.1177/000276426600900901.

Van Ingen, E., I. Stoop, and K. Breedveld. 2008. "Nonresponse in the Dutch Time Use Survey: Strategies for Response Enhancement and Bias Reduction." *Field Methods* 21(1): 69–90. DOI: https://doi.org/10.1177/1525822X08323099.

Van Kenhove, P., K. Wijnen, and K. de Wulf. 2000. "The Influence of Topic Involvement on Mail-Survey Response Behavior." *Psychology & Marketing* 19(3): 293–301. DOI: https://doi.org/10.1002/mar.1053.

Wallace, D. 1954. "A Case For- and Against- Mail Questionnaires." *Public Opinion Quarterly* 18(1): 40–52. DOI: https://doi.org/10.1086/266484.

Zuzanek, J. 1998. "Non-Response in Time-Use Surveys: Do the Two Ends Meet?" *Loisir et Société / Society and Leisure* 21(2): 547–549. DOI: https://doi.org/10.1080/07053436.1998.10753668.

Received June 2019
Revised April 2020
Accepted August 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 907–931, http://dx.doi.org/10.2478/JOS-2020-0043

Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context

James Wagner[1], *Brady T. West*[1], *Michael R. Elliott*[1], *and Stephanie Coffey*[2]

Responsive survey designs rely upon incoming data from the field data collection to optimize cost and quality tradeoffs. In order to make these decisions in real-time, survey managers rely upon monitoring tools that generate proxy indicators for cost and quality. There is a developing literature on proxy indicators for the risk of nonresponse bias. However, there is very little research on proxy indicators for costs and almost none aimed at predicting costs under alternative design strategies. Predictions of survey costs and proxy error indicators can be used to optimize survey designs in real time. Using data from the National Survey of Family Growth, we evaluate alternative modeling strategies aimed at predicting survey costs (specifically, interviewer hours). The models include multilevel regression (with random interviewer effects) and Bayesian Additive Regression Trees (BART).

Key words: Survey cost models; machine learning.

1. Introduction

Surveys are conducted in an environment of uncertainty. Many key design parameters are random variables. This makes it difficult to optimize a survey design before the field period begins. A new approach to survey design, called responsive survey design, is a direct reaction to this uncertainty. In a responsive design, the estimates of key design parameters are updated during the field period based on incoming data. This allows surveys to "recalibrate" designs toward optimality.

Responsive survey designs were first proposed by Groves and Heeringa (2006). They proposed that such designs should include the following elements: a) the pre-identification of key design features that have the largest potential impact on costs and errors in the survey, b) identification of a set of indicators for the cost and error properties of those design features to be monitored during the field period, c) changes to the design

[1] University of Michigan, 4053 ISR, 426 Thompson St., Ann Arbor, MI 48104, U.S.A. Emails: jameswag@umich.edu, bwest@umich.edu and mrelliot@umich.edu
[2] U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746, U.S.A. Email: stephanie. coffey@census.gov

Acknowledgments: This work was supported by a grant from the National Institutes for Health (#1R01AG058599-01; PI: Wagner). The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC's) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or the other funding agencies.

features over periods of time known as phases, based upon pre-specified decision rules, and d) combining data from the separate phases into a single estimate. The goal is to optimize cost-error tradeoffs over the phases. These tradeoffs are made within a total survey error framework (Biemer et al. 2017), where the total error is minimized for a fixed budget. For example, an initial phase might be highly successful at recruiting women but much less successful at recruiting men. A subsequent phase might be designed to complement this initial phase. The follow-on phase would aim to increase the participation of men.

Key to this process is monitoring the incoming data to know when a phase has reached its capacity and is no longer effectively reducing survey error. This might happen when a recruitment protocol is creating imbalances in who responds, or when the protocol becomes ineffective and very costly. The latter might occur, for example, in a face-to-face survey when continued contact attempts lose their effectiveness and become costly relative to the return.

The literature presents evidence of the successes that are possible when applying Responsive Survey Design (RSD) in practice. For example, the implementation of RSD in the National Survey of Family Growth led to a 25% reduction in the per-interview cost (Kirgis and Lepkowski 2013). Other projects have also implemented responsive survey designs with demonstrated success (Mohl and Laflamme 2007; Peytchev et al. 2009; Tabuchi et al. 2009; Kleven et al. 2010; Laflamme and Karaganis 2010; Lundquist and Särndal 2013; Barber et al. 2011; Finamore et al. 2013). However, these successes have been small in number and limited in scope (for a review, see Tourangeau et al. 2017).

One reason for the limited success may be that predictions about future outcomes (survey variables and costs) under alternative designs may be inaccurate. These sorts of predictions may be used to determine when a phase is over (Rao et al. 2008; Wagner and Raghunathan 2010; Lewis 2017; Paiva and Reiter 2017) and what the design of the next phase should be (Luiten and Schouten 2013; Rosen et al. 2014; Lynn 2016; Plewis and Shlomo 2017; Durrant et al. 2017). In either instance, the effectiveness of the responsive design is likely to be reduced if the predictions are inaccurate. Inaccurate predictions could lead, for example, to placing a phase boundary too early with the result that a more expensive design is implemented before the less expensive design has been fully exhausted. Inaccurate predictions about the types of respondents likely to be recruited over the different phases can lead to inefficiencies. It is also true that inaccurate predictions about the costs of a new phase could also lead to inefficient decisions. To date, no studies have attempted to evaluate methods for predicting costs in a responsive design framework.

This study attempts to address this gap by developing methods for predicting costs. These cost predictions would be a useful input for comparing alternatives and making design decisions. In this study, we assess our ability to accurately predict costs using different modeling approaches. We begin with a description in the Background section of the role of survey costs in the responsive design framework. We then discuss in the Data section the survey we will be examining and the data we use from that survey for our analyses. We then evaluate three different approaches to predicting costs in a face-to-face survey. These approaches are described in the Methods section. The approaches involve using existing data to predict costs with three different modeling strategies. We then compare the results across replications from six different time points.

2. Background

Responsive survey designs optimize tradeoffs between survey errors and costs over a series of phases. Knowing about the likely errors and costs under each phase is an important aspect of an RSD. Groves and Heeringa (2006) specify that survey designers should choose a set of indicators for both costs and errors that will be monitored by the survey during data collection. These indicators are then used as inputs to the responsive design. Specifically, these indicators are used as inputs to decision rules about when to change design phases and what the design of the next phase should be.

Some studies have examined how to specify a rule for determining when the current design is not leading to changes in estimates (Rao et al. 2008; Wagner and Raghunathan 2010; Lewis 2017; Paiva and Reiter 2017). Other studies have examined inputs to decision rules, such as predicted response propensities (Luiten and Schouten 2013; Rosen et al. 2014; Lynn 2016; Plewis and Shlomo 2017; Durrant et al. 2017). To date, no studies have examined the most effective methods for the real-time prediction of costs. Groves and Heeringa (2006) report on the impact of responsive design on survey costs, but in a *post hoc* analysis of the costs. Other studies have focused on the error side of the cost-error tradeoff (see Groves 2006 for a review of studies on nonresponse bias; see also Biemer and Trewin 1997 for a review of studies on measurement error). In fact, the study of survey costs, in general, is limited. In particular, we are aware of no studies focusing on the prediction of survey costs.

Improving the predictions of costs and errors may be key to improving responsive survey designs. A recent review of responsive and adaptive designs found that the impact of these designs was often limited (Tourangeau et al. 2017). One issue that may blunt the impact of these designs is inaccurate predictions of future costs or errors. These inaccuracies could lead to less than optimal designs, thereby mitigating the impact of interventions. For example, an error in the prediction of the impact of a changed incentive may lead to a design change that occurs too late, thereby prolonging a relatively inefficient design and reducing the overall efficiency of the survey.

Techniques for improving predictions are a burgeoning area of research. In addition to standard regression techniques, machine learning techniques have been applied to prediction problems in many fields, but less so in survey research (for a review, see Kern et al. 2019). Exceptions include using machine learning to code open-ended responses (Schonlau and Couper 2016). In particular, a class of models known as Bayesian Additive Regression Trees (BART, Chipman et al. 2010) may be useful for the prediction of survey costs. BART models have been used in a variety of settings. For example, they have been used to detect spam (Abu-Nimeh et al. 2008), model treatment effects in an experiment with survey questionnaires (Green and Kern 2012), predict driving behavior (Tan et al. 2018), and inform survival analysis (Sparapani et al. 2016). These models can also be used for both continuous and binary outcomes. In this study, we evaluate the ability of BART models to improve predictions of survey costs (specifically, interviewer hours expended) relative to regression modeling in a responsive design framework.

3. Data

3.1. Description of the Survey

The data for the present study come from the National Survey of Family Growth 2011–2019 (NSFG). This survey collects information on family formation, fertility, and other related topics. The survey population is persons living in the United States between the ages of 15 and 49 (prior to September 2015, the eligible population was persons in the United States between the ages of 15 and 44). A complete description of the survey, including questionnaires and field procedures, is available at http://www.cdc.gov/nchs/nsfg/.

The NSFG is a face-to-face survey that is conducted continuously. There are four quarters of data collection each year. In this analysis, we use data from 27 quarters (dating back to September 2011), but make predictions for six quarters (approximately January 2017–June 2018). The NSFG has a multi-stage area probability sample design. The primary stage units (PSUs) are counties and Metropolitan Statistical Areas. Each year, a new sample of PSUs is released. The second stage units (SSUs) are neighborhoods defined by Census Blocks. Each quarter, a new sample of SSUs is released within each of the sample PSUs. A sample of housing units is then released within each PSU. Interviewing is conducted in two stages. Interviewing staff first attempt to visit each housing unit and determine whether an eligible person lives there. This is known as “screening.” Once an eligible person has been identified and selected, an in-depth interview on fertility, family formation, and related topics is attempted. This second stage is known as the “main” interview.

The NSFG uses a responsive design approach. There are two phases. The first phase is defined by time and is completed in exactly ten weeks. The first phase design includes prenotification by standard mail, no token of appreciation for the screening interview, and a promised USD 40 token of appreciation for completion of the survey. The ten-week phase boundary was determined to be optimal using data from prior experience with NSFG data collection (Kirgis and Lepkowski 2013) and is fixed in advance.

The second phase lasts two weeks. In the second phase, a subsample of active cases is selected, and prenotification occurs using a Priority Mail package. This prenotification includes a token of appreciation for the screening interview (USD 5) for households that have yet to complete that stage and a prepaid token of appreciation (USD 40) for households where screening has been completed and an eligible person has been selected, with an additional token (USD 40, or USD 80 total) promised. This design has been shown to be effective in reducing the bias of NSFG estimates (Peytchev et al. 2010; Axinn et al. 2011).

In this design, the phase boundary is fixed (at ten weeks). In a responsive survey design, the phase boundary should be determined by observed changes in the field. The decision would be triggered by proxy indicators for costs and survey errors. In this study, our objective is to evaluate the ability of different methods to predict the costs associated with phase two of this design. The phase two costs include the cost of the mailed materials, the incentives, and the cost of interviewer time. If we can effectively predict the costs associated with this design phase, these predictions could be used to determine a more optimal time to switch to the second phase.

3.2. Description of Data

The data for this study are drawn from the following sources:

1. *NSFG sampling frame*. The sampling frame includes U.S. Census data on area characteristics and commercially supplied data on a large proportion of households.
2. *Paradata*. The NSFG paradata include interviewer observations about sampled neighborhoods, housing units, and persons, and level-of-effort data (e.g., number of call attempts with different types of outcomes, number of trips, etc.).
3. *Interviewer timesheet reports*. These data include the number of hours worked each day.

In this section, we describe briefly the variables drawn from each source and how they are summarized to the interviewer-week level in order to be used in the models. A full description of each of the variables is given in Appendix 1 (Subsection 7.1). The goal of the study is to compare the ability of different modeling approaches to predict survey costs. Therefore, we intend to use predictors that are available at the time that those predictions need to be made. In this case, that is two weeks prior to the week for which the predictions are to be made. In week ten of an NSFG quarter, predictions for weeks 11 and 12 (the weeks defining the second phase) would be needed in order to make a decision about whether to change the design and use the second phase protocol.

For this study, the main cost driver is interviewer hours expended. The number of hours depends, in part, upon the phase, as the second phase involves subsampling cases. There are other costs associated with the change in phase. These are the special mailing and the additional tokens of appreciation for the screening interview and the main interview. We do not predict these costs using the models. The sample size for the mailing is known prior to the phase. The costs of the increased incentive, which is prepaid, are also known. The cost of the post-paid incentive is a function of the number of interviews. We use response propensity models (described elsewhere, see West et al. 2019) as the basis of estimating these costs. Our primary focus in this study is on prediction of the hours that interviewers will expend once the second phase begins.

The sampling frame data include geographic characteristics, such as the Census Division (the United States is divided by the U.S. Census Bureau into nine Divisions), as well as area characteristics, such as estimated eligibility rates for the Census Tract (a geography defined by the U.S. Census Bureau containing roughly 2,500 to 8,000 persons) of the sampled unit from the American Community Survey (ACS) and estimated rates for the U.S. Census Block Group (a geography defined by the U.S. Census Bureau usually corresponding to between 600 and 3,000 persons) of ever being married from the ACS. The urbanicity of the sampled area is assigned at the county level using the U.S. Office of Management and Budget's (OMB) definition of Metropolitan Statistical Areas. Also available on the sampling frame are data from a commercial vendor, such as the ages of adults in the household. These data are linked to sample addresses in advance. There are a proportion of households for which these data are not available, and they may also be inaccurate (see West et al. 2015 for an appraisal). The variables include the estimated age of one or two persons in the household, the estimated household income, and the quality of the match (i.e., likelihood of the match being correct).

These sampling frame data are attached to NSFG call record data. These predictors are summarized up to the interviewer-week level, either by taking the mean (for continuous variables, e.g., rates of ever being married at the U.S. Census Block Group level) or the mode (for categorical variables, e.g., Census Region – 50 States and the District of Columbia are grouped into four Census Regions, and nine Divisions nested with Region) of these characteristics for all the contact attempts each interviewer made in a given week. Since these contact attempt data are not available for "future" weeks, we lagged these variables by two weeks in order to make them available for prediction of the costs two weeks into the future. For example, we use the modal urbanicity of cases that were attempted two weeks prior as a predictor of interviewer hours in the current week. This variable could be observed in week ten in order to predict hours for weeks 11 and 12.

Paradata used in this study include interviewer observations of neighborhood characteristics, such as whether there are unimproved roads or seasonal hazards that make access to the neighborhood difficult and whether there is evidence of speakers of languages other than English. Interviewers also observe characteristics of housing units, such as whether it is a single-family home or a multi-unit structure, whether there is evidence of children in the household, and the likelihood that all persons in the household are over the age of 45. There are also level-of-effort paradata variables that are derived from records of call attempts and information about the number of active lines.

The paradata from any given week are highly predictive of the hours of effort in the same week (Wagner 2019). However, in this case, we are making predictions for the future and only have the values for these variables from previous time periods. Therefore, we included a series of variables that summarize counts of call attempts with different result types (e.g., main interviews, screening interviews, setting appointments) for the time period two weeks prior to the weeks for which predictions are being made. Since many of the predictors are lagged values, the first two weeks of each quarter were excluded from the analysis. We also have both the number of area segments visited and the number of active sample housing units in each interviewer's workload two weeks prior to the week for which predictions are being made. These variables are available at the point in time when cost predictions are going to be made. We also include an indicator for whether the week is in the second phase of the NSFG responsive design. The phase variable could be used to predict costs under two different design options – the phase one design versus the phase two design.

The cost data are derived from interviewer timesheets. There are three predictors derived from the timesheet data. The first is the number of hours worked in the week that was two weeks prior to the week for which predictions are being made. For example, the hours worked in week eight will be used to predict the number of hours worked in week ten. A second predictor is a categorical variable indicating whether the interviewer was involved with overnight travel two weeks ago. This measure is reported as the number of hours each week that are spent in overnight travel, but is categorized here as none, some, or all of the hours in the week two weeks prior to the current week. A third variable, days worked, is a count of the number of days in a week (two weeks prior to the current week) that the interviewer reported any time on their timesheet.

Our primary dependent variable is the hours worked by the interviewer each week, which is set by the interviewer in consultation with their manager. NSFG interviewers agree to work a minimum of either 20 or 30 hours per week. However, interviewers set their own schedules and may deviate from the minimum specified for various reasons – either due to requests from sampled persons for specific appointment times, low contact rates on a particular day (Wagner and Olson 2018), or for personal reasons. Therefore, the hours worked for each interviewer is a random variable.

Each record in the dataset represents a week of an interviewer. In other words, an interviewer who works all weeks in a 12-week quarter will have ten records in the dataset (excluding weeks one and two due to the use of lagged values). Each interviewer is scheduled to work for at least one year (four quarters). This doesn't always happen (interviewers sometimes resign mid-quarter) and some interviewers worked in multiple years. In the data, the mean number of records (weeks) per interviewer was 47.6. The median was 27. The minimum and maximum were one and 253, respectively. Each record includes the number of hours worked that week (the outcome), a set of known characteristics (the week, the phase, the quarter, and the year), hours worked two weeks ago, whether the interviewer was travelling overnight two weeks prior, summary information about call attempts made two weeks prior, and characteristics of the interviewer's sample two weeks prior. Some records were excluded from the analysis. In some weeks, an interviewer might have reported hours in their timesheet, but did not make any call attempts. On rare occasions, call attempts were recorded in a week for which no hours were reported. In both of these situations, these records were excluded from the analysis.

4. Methods

4.1. Alternative Modeling Approaches

We will examine three different modeling approaches to predict the interviewer hours in a given week: a linear mixed (multilevel) model (MLM) with random intercepts for each interviewer, a Bayesian Additive Regression Trees (BART) model (Chipman et al. 2010) that includes an indicator variable for each interviewer, and a random intercept BART (RI BART) model (Tan et al. 2018). We also attempted to fit a linear regression model that did not include an indicator for each interviewer. This model produced poor predictions and, therefore, was not included.

From previous research into survey errors, we know that interviewers vary in their ability to recruit participants (see West and Blom 2017, for a review). Hence, we use multilevel models (MLM) in which each interviewer has a random intercept:

$$y_i = \sum\nolimits_{p=1}^{P} x_{ip}\beta_p + \alpha_{g[i]} + \varepsilon_i$$

where g indexes the interviewer who may have repeated (i.e., clustered) measurements and p indexes the set of covariates, including a vector of 1's. The "random intercepts" associated with each interviewer are assumed to be draws from a normal distribution with a common variance. This allows for random variation among interviewers to be incorporated into the model. These models assume a linear relationship between the predictors and the outcome. In this case, that seems reasonable as the coefficients represent estimated changes in hours when characteristics of the sample are changed. This method has been used previously to estimate costs in surveys (Wagner et al. 2017).

Our second method uses Bayesian Additive Regression Trees with an indicator variable for each interviewer. The BART approach uses the sum of a large number of regression trees, where each tree is constrained by specification of priors to include relatively few predictors. The model has a very general form:

$$y_i = \sum_{j=1}^{m} f(\boldsymbol{X}_i; T_j, M_j) + \varepsilon_i$$

where *f*(.) defines the BART model, and *T* denotes a sequence of decision rules that split the sample into groups with each terminal node having a mean μ_i from the set *M*. The process can be repeated over *m* trees (indexed by *j*), with each tree reducing $\sum_{i=1}^{n} \varepsilon_i$ from the previous trees. Posterior distributions on parameters are developed using a Markov Chain Monte Carlo (MCMC) algorithmic approach. BART models have the ability to include or exclude a large number of interactions, including polynomials of continuous predictors, as the data will suggest. BART model predictions have been shown to perform well against other machine learning techniques (Chipman et al. 2010), and they allow for the calculation of a principled measure of uncertainty associated with the predictions. We would expect the BART model to produce more accurate predictions than MLM (Chipman et al. 2010). However, BART models do not produce easily interpretable model estimates in the same way that regression models produce estimated coefficients. Instead, BART models produce predicted outcomes.

We estimated BART models using one with interviewers as fixed effects and another incorporating interviewer effects as random intercepts (RI BART; Tan et al. 2018). The latter model can be specified as:

$$y_i = \sum_{j=1}^{m} f(\boldsymbol{X}_i; T_j, M_j) + \alpha_{g[i]} + \varepsilon_i$$

where we use the same notation as earlier, with the addition of the random intercept, $\alpha_{g[i]}$, and *g* indexes the interviewer, who may work multiple weeks and therefore produces a clustered set of observations. Both the MLM and RI BART models allow for some consistency within each interviewer in their expected hours each week; the random effect approach formally treats the interviewers as being drawn from a population of potential interviewers, and can stabilize estimates of these "interviewer effects'' compared with treating them as fixed effects.

We made predictions of hours worked during phase two for each of six quarters, where we trained the models using data from previous quarters and the first phase of the current quarter to predict the hours expended during the second phase of the current quarter. For example, we predicted the hours worked for phase two of Q22 using the hours from Q1 through Q22 phase one. We made predictions for phase two hours for Q22 through Q27. This allows us to assess the performance of our models using a form of temporal cross-validation. Given that we know the actual hours worked in these six quarters, we evaluate the predictions using two measures of prediction accuracy: mean squared error (MSE) and mean absolute error (MAE). We compare the observed and predicted values for our dependent variable (interviewer hours expended) in the two weeks of phase two across six quarters of data collection.

All models included the predictors listed in Appendix 1. The multilevel models included a random intercept; the BART models included a dummy variable for each interviewer to accommodate interviewer effects if they were present; and the RI BART models included a random intercept for each interviewer.

All models were fit in R. The multilevel model was fit using the `lmer` function in the lmer4 package. The BART model was fit using the `bart` function in the dbarts package (Dorie et al. 2019). In these BART models, the interviewer ID was included as a factor, so that an indicator for each interviewer could be included in the models. The RI BART model included a random intercept for each interviewer. These models were fit using the `rbart_vi` function in the dbarts package.

The BART models require the specification of priors on several parameters. In many cases, the default settings will work quite well (Chipman et al. 2010). In our case, in order to determine the values for priors, we performed cross-validation. The actual analyses were run on quarters 22 to 27. Therefore, in order to set priors, we tested a range of priors for quarters 16 to 21 and used training and test samples to determine the MSE and MAE for each combination of priors tested. Two important priors are the number of trees and k (the number of standard deviations $E(Y|x) = f(x)$ is away from +/-.5). We tested a grid of possible values for trees of 45, 100, 150, 200, 250, 300, 350, and 400 crossed with the possible values for k of 1, 2, 3, and 4. We tested this grid of combinations of values across both the BART models. We calculated the MSE on the test sample for each of the six quarters (15 to 21) and then ranked the parameter pairs with lowest to highest MSE. In the case of the BART models, we found that 300 trees and k=1 performed the best with a mean rank of 4.0 across the six quarters. For the RI BART models, we found that 300 trees and k=2 performed the best with a mean rank of 4.0.

Once we had selected these two parameters for each of the two BART models, we also tested values for the priors for the variance parameters in a similar manner. The two parameters are the degrees of freedom for the error variance and the quantile of the error variance. We chose as the prior for the degrees of freedom for the variance as 3.1 and the quantile of the error variance prior as 0.96. These values are close to the suggested default values of 3.0 and 0.90.

Finally, for the BART models, we needed to set the MCMC parameters. We identified parameters that worked well for all quarters, and then ran the models using those parameters. We used trace plots of key parameters and a review of plots of autocorrelation functions to monitor the MCMC. In the case of the BART models, that meant 500 iterations for a burn-in, thinning to one in every 3,200 iterations, and running until 1,000 draws were obtained. For the RI BART models, we had 500 burn-in iterations, thinning to one in every 350 iterations, and running until we had 1,142 draws. We ran four chains for a total of 4,568 draws.

5. Results

First, we examine which predictors were important in each model. In the multilevel regression model, the interviewer IDs were important predictors. The proportion of the total variance in costs that is due to the interviewers was between 0.21 and 0.25 across the six quarters we considered. These intra-class correlations are quite substantial relative to those for other survey outcomes (West and Blom 2017), suggesting that NSFG interviewers vary substantially in terms of these weekly hours during the second phase. Full model results for one quarter (Q27), as an example, are available in Appendix 2 (Subsection 7.2).

The BART models do not create scoring functions in the same way that a regression model does. Instead of presenting coefficients, we display the 20 predictors used most frequently in the BART modeling process. For the BART model, Figure 1 shows the 20 predictors that were used most frequently in splits in the Q27 model predicting the hours worked during an interviewer week. This is calculated as the number of splits based on that variable, divided by the total number of splits used. This quantity is averaged across all of the draws. Each predictor is described in more detail in Appendix 1 (Subsection 7.1).

The predictor most frequently included in the models is the number of hours worked in the week two weeks prior to the week of interest. The next most frequently used variable is an indicator variable for interviewer #78. In fact, five of the top ten variables are indicator variables for interviewers. Similar to the results from the multilevel model, this suggests that interviewers are relatively consistent in the hours that they charge each week. The number of active sample lines two weeks prior is the third most frequently used predictor for splits. An indicator variable for whether the interviewer was on full time travel two weeks prior to the week of interest is also important, as is an indicator for phase two. Figure 2 presents a similar figure for the RI BART model for Q27.

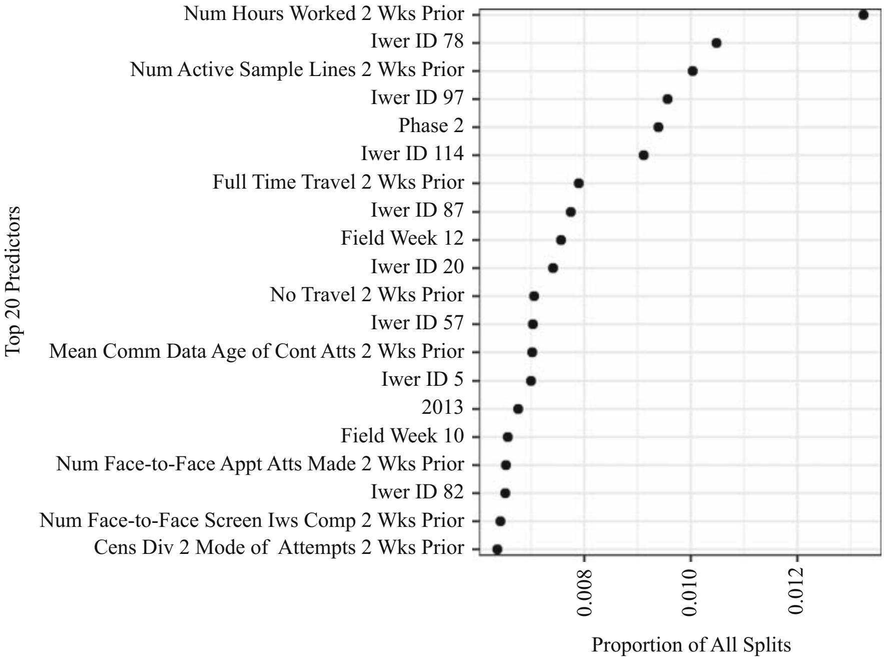


Fig. 1. The 20 predictors with the highest proportion of splits based on the variable in the ensemble of trees in the BART model for Q27 costs.

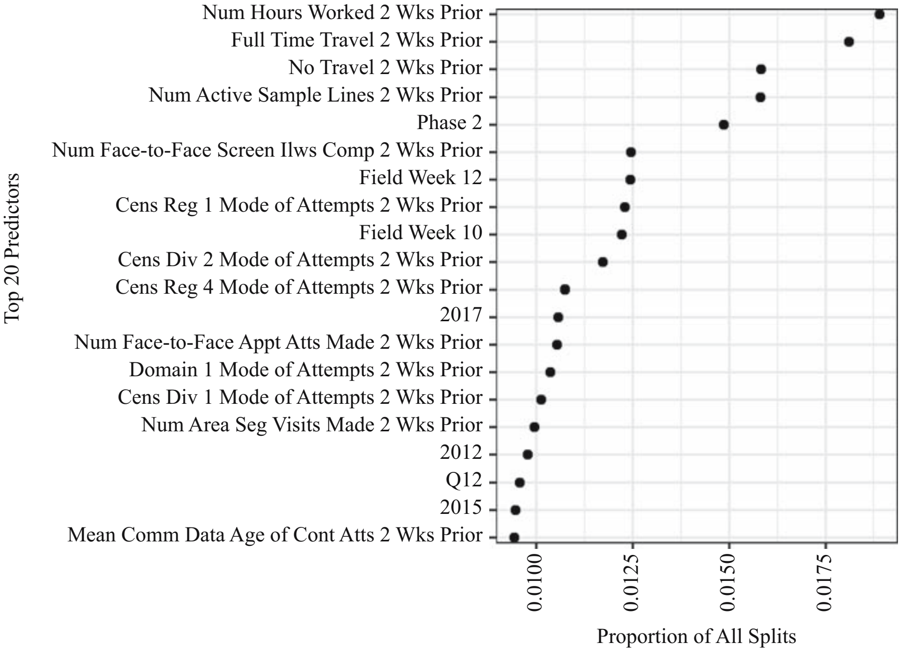


Fig. 2. The 20 predictors with the highest proportion of splits based on the variable in the ensemble of trees in the RI BART model for Q27 costs.

The number of hours worked two weeks prior to the current week is the most frequently included variable. Whether the interviewer was on full time travel or, alternatively, did not travel at all two weeks prior to the target week were also important predictors. An indicator for phase two was once again an important predictor. The number of screening interviews completed two weeks prior is also an important predictor. Other important predictors are indicators for several of the Census Divisions or Regions, sampling domain 1, several years (2012, 2015, and 2017), an indicator variable for quarter 12 (Q12), and the number of area segments visited and the number of appointments made two weeks prior to the current week.

Next, we examine the predictions of overall costs. Figure 3 presents the predicted total interviewing hours in phase two for each of the six "testing" quarters from each of the two modeling approaches. For each quarter, the predictions are based upon the data observed prior to each quarter's phase two. The predictions of the total also have error bars. For the multilevel models, these are 95% bootstrap confidence intervals. These are 95% credible intervals for the BART models.

The models produce predictions of the total phase two hours with similar quality. All models do well in Q22, Q25, Q26, and Q27. Notably, the BART and RI BART models produce consistently narrower credible intervals. However, in Q23 and Q24, none of the models perform particularly well, and the intervals do not include the total hours that were actually observed. However, the intervals surrounding the predictions of the RI BART models come closer to including the observed hours than do those for the BART or MLM models. In all of the quarters, the RI BART model provides the predicted total closest to the observed total.

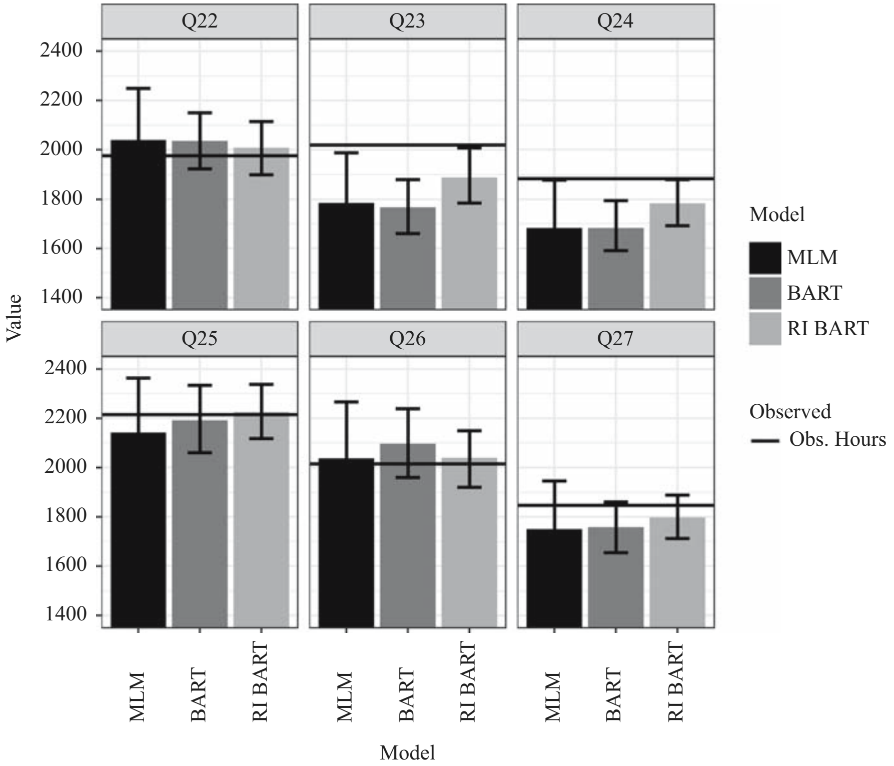


Fig. 3. Predicted total hours and 95% bootstrap or credible intervals from the MLM, BART, and RI BART models for each of six quarters.

Although the models predicted the total phase two costs fairly well, the differences between interviewers in pay rates mean that predictions at the interviewer level may be important in obtaining an accurate prediction of the overall cost. In Figure 4, we examine the predictions for each interviewer-week in phase two of Q27. Each dot in Figure 4

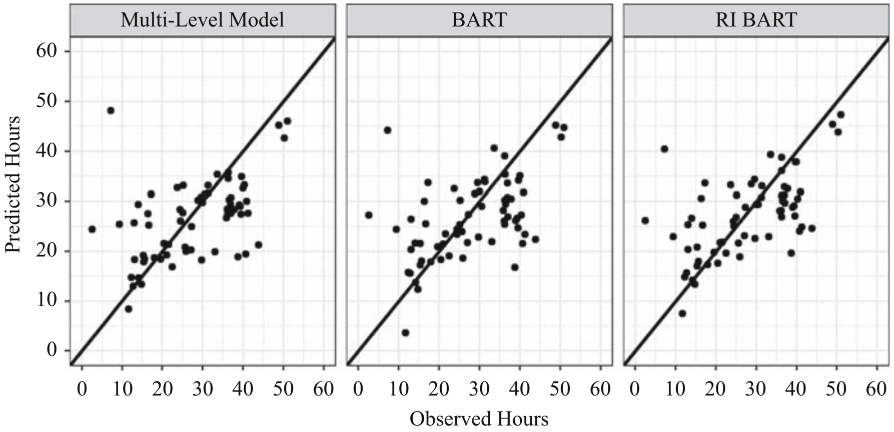


Fig. 4. Q27 hours: Predicted values from three different models versus observed values.

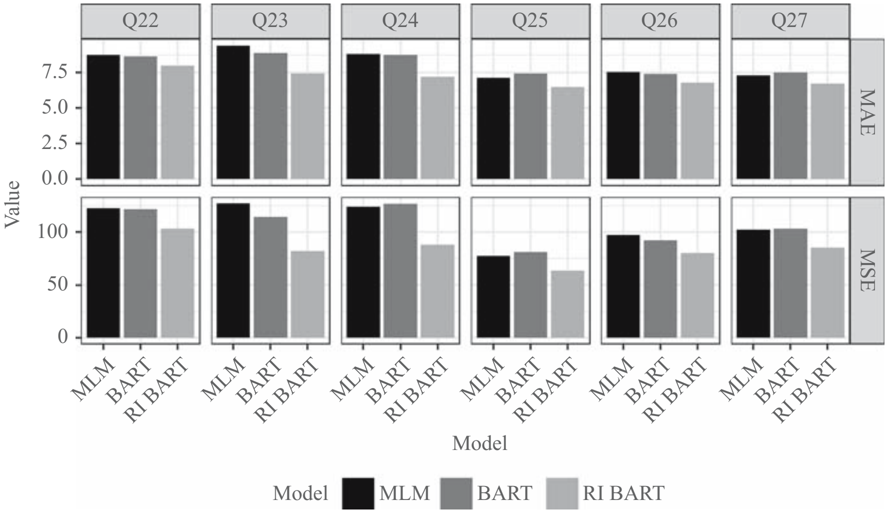


Fig. 5. Mean squared error (MSE) and mean absolute error (MAE) for three models (multi-level model, BART, and RI BART) across six quarters (Q22–Q27).

represents one week worked by an interviewer. The observed hours are on the x-axis and the predicted hours are on the y-axis. The 45-degree line represents agreement between the predicted and observed values.

The RI BART model does appear to provide better predictions. To confirm this, we summarize the results at the interviewer-week level for all quarters using two error measures – the mean squared error and the mean absolute error. Figure 5 presents these measures for each of the six quarters.

The results in Figure 5 indicate that the RI BART model has the best performance for predicting interviewer-level costs. The RI BART model has the lowest MSE – and often substantially lower – than the other models in all six of the quarters. The BART model, on the other hand, produces predictions that have the second lowest MSE in three of the quarters (Q22, Q23, and Q26). The MLM models have the second lowest MSE in the other three quarters (Q24, Q25, and Q27).

6. Discussion

The accurate prediction of expected costs associated with a design change in a responsive survey design framework is necessary in order to make cost-error tradeoff decisions about potential design changes across phases of a survey. In our case, we are comparing two designs (phase one versus phase two). The predictions of costs – along with predictions about expected errors under different designs – could be used to optimize the design. In this case, it would be possible to use these predictions to determine when to switch from phase one to phase two. As a first step in this direction, we evaluate our ability to predict costs.

In some settings, the prediction of costs will be tightly tied to the completion rate. For example, in a web survey, the costs are largely driven by the payment of incentives and, to a lesser extent, the cost of sending email reminders. In that case, the predictions of costs are largely a function of expected completion rates under different designs. It would be possible to accurately predict costs based on different incentive amounts if an accurate prediction of the completion rate at each incentive amount is available.

In our setting, prediction of costs is more complicated. Interviewers make frequent scheduling changes. Often, this is done in response to requests from sampled units. Interviewers seek to accommodate the schedules of sampled persons. For example, interviewers will set appointments with sampled persons at times that are convenient for the sampled person. This might mean making a change to their planned schedule for the week. Interviewers may also shorten their scheduled hours when they experience relatively low contact rates (Wagner and Olson 2018). These accommodations to the interviewers' schedules are necessary given the need to be flexible with the schedules of sampled persons. Further, the level of effort needed to obtain an interview can vary greatly from interviewer to interviewer and week to week within an interviewer. This can be a function of the choices the interviewer makes about when to work, the characteristics of the sample, and other factors that appear as noise in models of response propensity. These factors make it difficult to predict the number of hours that an interviewer will work in any given week. There are, on the other hand, factors that stabilize interviewers' effort. Mainly, project managers seek stable effort over time. This might include a certain number of hours as a goal to be worked each week. The hours that interviewers work each week may also vary when design changes are introduced over time. Our analysis focused on predicting hours as a function of stable design features.

We found that we could develop relatively accurate predictions with relatively simple models. The multilevel and BART models generated accurate predictions of the total hours worked in a given week. However, differences in wages paid across interviewers mean that these estimates of total hours might give very different cost estimates. The model that produced the most accurate results was the random intercept BART ("RI BART") model. The BART model that treated interviewers as a fixed effect also performed well. In initial analyses, we found that using a BART model (not with random intercepts) with many of the default settings performed about as well as the multilevel models. In our setting, the ability to tune the parameters proved valuable. We were able to do this since we had 27 iterations of the same survey. In other settings, the inability to test a variety of parameters might reduce the effectiveness of the BART approach, although Chipman et al. (2010) found that the default settings perform well in several different situations.

This article has focused exclusively on the ability of different modeling approaches to improve the prediction of survey costs. These predictions are intended to inform decisions about interventions in a responsive survey design (RSD) context. Although it is clear that measures of cost are always relevant for making design decisions, the exact way in which these predictions would be used to make decisions during data collection was outside the scope of this article. Here, we briefly outline an approach to including these predictions in a decision framework for RSD.

In the specific case presented in this article, the decision is whether to switch the data collection protocol from phase one to phase two. In order to optimize that decision, we need to have estimates of the impact of phase two on nonresponse bias and the costs of the second phase. A further consideration would be sampling error. A simple form of the optimization problem would look at the design decision each week and compare outcomes for two design options under the same fixed budget: 1) continue with phase one, or 2) switch to phase two.

The outcome to be compared between the two designs would be the expected mean squared error of the survey estimates (incorporating both expected nonresponse bias and sampling error). Phase one could produce more interviews but risk higher nonresponse bias, while phase two could lower the risk of nonresponse bias but produce fewer interviews.

The focus of this article has been on accurate predictions of the costs involved in this decision. Other papers have focused on relative nonresponse errors associated with similar types of design changes (Peytchev et al. 2010; Axinn et al. 2011). With predictions of costs under the two approaches – which could be generated using the approach outlined in this article – and predictions of which cases are likely to be interviewed and their predicted responses under the two design options, the option that minimizes mean squared error for a fixed budget could be selected at a given point in time. The approach outlined here also allows for uncertainty in the cost estimation to be built into the decision-making process, for example, by making decisions based on a low or high percentile of an estimate, rather than just a mean or to make probability statements about predicted outcomes.

We also note that predictions have variance and accounting for this variance may be important. In our case, this variance was captured through repeated draws from the posterior distribution for the BART models. We used a bootstrap approach to assessing variance of the MLM predictions. We know that incorrect assumptions about underlying cost parameters can lead to inefficient designs (Burger et al. 2017). Capturing the variance may be a helpful input to design decisions. Designers, rather than focusing on point estimates, might use the variance to calculate the probability of achieving survey goals under alternative designs. Such an approach would lead to better decisions. The models developed in this paper and the predictions and prediction intervals they produce provide one "leg" of such a decision analysis.

Future research could look at combining predictions about costs, response propensities, and survey outcome variables under different designs. Then, design decisions can be informed by the predicted errors and costs under the alternative designs. This framework would allow us to move closer to a "total survey error" approach in practice. Of course, the underlying models need to be evaluated. This article is a step toward the development of models aimed at predicting costs. Many other models will need to be developed and tested for settings unlike the one used in this paper.

7. Appendix

7.1. Appendix 1. Available predictors used in all three types of models.

Source	Predictor	Description
	NEWIWERID	Interviewer ID Number
TIME-SHEETS	NHOURS_LAG2	Number of hours worked by the interviewer in the week two weeks prior to the current week
	TRAVEL_LAG2	How much did the interviewer participate in overnight travel in the week two weeks prior to the current week: NONE, SOME, or ALL of the week. Generally, interviewing staff is split into those who travel and

7.1. Appendix 1. Continued.

Source	Predictor	Description
		those who do not. However, sometimes, under special circumstances such as a need to infuse more hours into production due to an unplanned staff shortage or if a PSU in need of hours happens to be geographically close to another PSU, non-travelling interviewers will travel. Therefore, this variable has three categories.
	DAYS_ WORKED_ LAG2	The number of days worked (i.e., days with an entry in the timesheet) in the week two weeks prior to the current week
SAMPLING FRAME	QTR	The quarter of production (Q1–Q27)
	YEAR	The calendar year of production (2011–2018)
	CENSUS_DIV_ MODE_LAG2	The modal Census Division of the lines attempted by an interviewer in the week two weeks prior to the current week.
	CENS_REG_ MODE_ LAG2	The modal Census Region of the lines attempted by an interviewer in the week two weeks prior to the current week.
	EST_ELIG_ RATE_ MEAN_LAG2	This is the mean of the Census ZIP Code Tabulation Area (ZCTA) level data about the estimated eligibility rate. The data are at the ZCTA level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	EST_ELIG_ 15_49_ACS_ MEAN_LAG2	The mean of the estimated eligibility rate for the Census Block Group reported in the American Community Survey. The data are at the Block Group level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	ELIG_NEVER_ PCT_ MEAN_ LAG2	This is the percentage of eligible persons living in the Census Tract who have never been married. The data are at the Tract level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	OCC_RATE_ MEAN_ LAG2	This is the Census Block level occupancy rate from the 2010 Decennial Census. The data are at the Block level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	DOMAIN_ MODE_ LAG2	The domain is set at the Census Block Group level and assigned to housing units within each BG. All BGs are assigned to a domain based upon the following definitions: 1) < 10% of Block Group African-American and < 10% Hispanic, 2) > = 10% of Block Group African-American and < 10% Hispanic, 3) < 10% of Block Group African-American and > = 10% Hispanic, and 4) > = 10% of

7.1. Appendix 1. Continued.

Source	Predictor	Description
		Block Group African-American and > = 10% Hispanic. The mode is for the domain of the lines that are attempted in the week two weeks prior to the current week.
	URBAN_MODE_ LAG2	The mode of the urbanicity (assigned at the case level) of the attempts made during the week that is two weeks prior to the current week, where 1 = Major Metropolitan Area, 2 = Minor Metropolitan Area, 3 = Non-Metropolitan Area, 4 = Remote Area.
INTER-VIEWER OBSER-VATIONS	STRUCTURE_ TYPE_ MODE _LAG2	The mode of the structure type variable of the cases that were attempted in the week that is two weeks prior to the current week. 1 = Single family home, 2 = Structure with 2 to 9 units, 3 = Structure with 10 + units, 4 = Mobile home, 5 = Other.
	BLACCESS_ GATED_ MEAN_LAG2	The mean of an area segment-level observation about whether there is a gated community in the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_ SEASONAL_ HAZARD_ MEAN_ LAG2	The mean of an area segment-level observation about whether there is a potential seasonal hazard preventing access to the area segment (e.g., unplowed roads). This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_ UNIMPROVED_ ROADS_MEAN_ LAG2	The mean of an area segment-level observation about whether there are unimproved roads limiting access to the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_ OTHER_MEAN	The mean of an area segment-level observation about whether there other (i.e., not gated, seasonal hazards, or unimproved roads) factors limiting access to the area segment. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	LRESIDENTIAL_ MEAN	The mean of an area segment-level observation about whether the area is completely residential or also includes some commercial structures. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	INON_ENGLISH_ SPEAKERS_ MEAN_LAG2	The mean of an area segment-level observation about whether the area has evidence of non-English speakers. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLNON_ ENGLISH_ LANG_SPANIS_ MEAN_LAG2	The mean of an area segment-level observation about whether the area has evidence of Spanish speakers. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ISAFETY_ CONCERNS_ MEAN_LAG2	The mean of an area segment-level observation about whether the interviewer had concerns about their safety on the first visit. This is observed at the segment level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	MANYUNITS_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the sampled housing unit has 1 = more than one unit, or 0 = 1 unit. This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	CHILDREN UNDER15_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the interviewer believes that there are children under the age of 15 living in the housing unit (1 = Yes, 0 = No). This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ALLAGE-OVER45_ MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the interviewer believes that persons living in the housing unit are all over the age of 45 (1 = Yes, 0 = No). This is observed at the housing unit level, but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
COMMERCIAL DATA	MSG_ MATCHQUALITY_ MEAN_LAG2	A variable indicating the estimated quality of the match of commercially-available data to the address (1–5). The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	MSG_AGE_ MEAN_LAG2	The mean age of the first person from the commercially-available data where those data are available. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	MSG_INCOME_ MEAN_LAG2	The mean of the estimated household income for cases with a match to commercially-available data. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
LEVEL OF EFFORT PARA-DATA	PHASE	The phase of the NSFG design (first phase occurs in weeks 1–10, phase two during weeks 11–12).
	LAG2.ACTIVE_ LINES	The number of active lines from two weeks prior to the current week for each interviewer.
	TRIPS_LAG2	The total number of unique visits to an area segment (derived from call record data) from two weeks prior to the current week for each interviewer.
	FTFNOCONTACT_LAG2	The total number of Face-to-face contact attempts that resulted in no contact from two weeks prior to the current week for each interviewer.
	FTFCONTACT_ LAG2	The total number of Face-to-face contact attempts that resulted in a contact with only agreement for a general callback from two weeks prior to the current week for each interviewer.
	FTFAPPT_ LAG2	The total number of Face-to-face contact attempts that resulted in setting an appointment from two weeks prior to the current week for each interviewer.
	MAINIW_LAG2	The total number of main interviews (all main interviews are completed face-to-face) from two weeks prior to the current week for each interviewer.
	FTFMAINCONCERN_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person expressing concerns from two weeks prior to the current week for each interviewer.
	FTFMAINNI_ LAG2	The total number of Face-to-face contact attempts that resulted in a final noninterview from two weeks prior to the current week for each interviewer.
	FTFMAINNS_ LAG2	The total number of Face-to-face contact attempts that resulted in a final nonsample from two weeks prior to the current week for each interviewer.
	FTFSCRNIW_ LAG2	The total number of Face-to-face contact attempts that resulted in a screening interview from two weeks prior to the current week for each interviewer.
	FTFSCRNCONCERN_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit expressing concerns prior to completing a screening interview from two weeks prior to the current week for each interviewer.

7.1. Appendix 1. Continued.

Source	Predictor	Description
	FTFSCRNNI_ LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as a noninterview prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	FTFSCRNNS_ LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as nonsample prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	FTF_MAINNS _INEL_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person being finalized as ineligible prior to completing a screening interview from two weeks prior to the current week for each interviewer.
	ACTIVE_ LINES_LAG2	The number of active sampled units two weeks prior to the current week for each interviewer.
	TEL_ALL_ LAG2	The total number of telephone attempts made by each interviewer two weeks prior to the current week.

7.2. Appendix 2. Q27 multilevel model predicting hours worked by an interviewer in a week: Estimated coefficients, confidence interval, and p-value.

Predictors	Estimates	CI	p
(Intercept)	32.73	24.32 – 41.13	**< 0.001**
Q10	-3.32	-5.02 – -1.62	**<0.001**
Q11	-1.60	-3.31 – 0.12	0.068
Q12	-0.23	-1.95 – 1.50	0.798
Q13	-2.71	-4.39 – -1.03	**0.002**
Q14	-1.83	-3.52 – -0.15	**0.033**
Q15	-1.26	-2.95 – 0.44	0.146
Q16	-1.40	-3.11 – 0.31	0.109
Q17	-1.29	-2.98 – 0.41	0.137
Q18	-1.05	-2.77 – 0.68	0.235
Q19	-2.34	-4.08 – -0.60	**0.008**
Q2	-3.37	-4.95 – -1.79	**<0.001**
Q20	-2.42	-4.16 – -0.67	**0.007**
Q21	-2.85	-4.68 – -1.02	**0.002**
Q22	-2.04	-3.82 – -0.25	**0.025**
Q23	-2.76	-4.56 – -0.95	**0.003**
Q24	-0.65	-2.47 – 1.17	0.483

7.2. Appendix 2. Continued.

Predictors	**Estimates**	**CI**	**p**
Q25	-2.30	-4.01 – -0.59	**0.008**
Q26	-2.74	-4.54 – -0.93	**0.003**
Q27	-3.35	-5.18 – -1.52	**<0.001**
Q3	-3.25	-4.86 – -1.64	**<0.001**
Q4	-0.95	-2.58 – 0.69	0.257
Q5	-2.46	-4.15 – -0.76	**0.004**
Q6	-2.45	-4.15 – -0.76	**0.005**
Q7	-3.33	-4.98 – -1.68	**<0.001**
Q8	-2.60	-4.27 – -0.92	**0.002**
Q9	-2.92	-4.59 – -1.25	**0.001**
PHASE_MODE2	-1.44	-2.05 – -0.83	**<0.001**
NHOURS_LAG2	0.13	0.09 – 0.17	**< 0.001**
DAYS_WORKED_LAG2	-0.10	-0.33 – 0.14	0.414
NONE	-0.84	-1.66 – -0.01	**0.046**
SOME	-1.31	-2.50 – -0.11	**0.033**
CENSUS_DIV_MODE_LAG22	0.74	-1.12 – 2.61	0.435
CENSUS_DIV_MODE_LAG23	0.06	-1.66 – 1.79	0.942
CENSUS_DIV_MODE_LAG24	-1.87	-3.92 – 0.18	0.074
CENSUS_DIV_MODE_LAG25	0.93	-1.17 – 3.03	0.384
CENSUS_DIV_MODE_LAG26	1.74	-0.66 – 4.15	0.155
CENSUS_DIV_MODE_LAG27	1.67	-0.78 – 4.12	0.181
CENSUS_DIV_MODE_LAG28	-0.52	-3.11 – 2.07	0.694
CENSUS_DIV_MODE_LAG29	0.95	-1.48 – 3.38	0.444
DOMAIN2_MODE_LAG22	0.65	-0.07 – 1.38	0.078
DOMAIN2_MODE_LAG23	-0.13	-0.92 – 0.65	0.738
DOMAIN2_MODE_LAG24	0.35	-0.47 – 1.17	0.406
URBAN_MODE_LAG22	-0.31	-1.11 – 0.48	0.436
URBAN_MODE_LAG23	0.23	-1.21 – 1.66	0.758
URBAN_MODE_LAG24	5.73	-1.50 – 12.96	0.121
STRUCTURE_TYPE_MODE_LAG22	-0.21	-1.17 – 0.75	0.663
STRUCTURE_TYPE_MODE_LAG23	-0.17	-1.13 – 0.78	0.723
STRUCTURE_TYPE_MODE_LAG24	0.98	-0.71 – 2.67	0.256
STRUCTURE_TYPE_MODE_LAG25	-5.87	-25.47 – 13.73	0.557
BLACCESS_GATED_MEAN_LAG2	0.09	-0.54 – 0.71	0.782
BLACCESS_SEAS_HAZARD_MEAN_LAG2	0.59	-0.61 – 1.80	0.335
BLACCESS_UNIMP_ROADS_MEAN_LAG2	-0.69	-1.60 – 0.21	0.133
BLACCESS_OTHER_MEAN_LAG2	-0.21	-1.33 – 0.90	0.709
LRESIDENTIAL_MEAN_LAG2	0.00	-0.65 – 0.66	0.988
INON_ENGLISH_SPEAKERS_MEAN_LAG2	-0.28	-0.89 – 0.33	0.372
ISAFETY_CONCERNS_MEAN_LAG2	0.73	0.11 – 1.35	**0.020**
MANYUNITS_MEAN_LAG2	-0.20	-1.63 – 1.23	0.785
CHILDRENUNDER15_MEAN_LAG2	0.42	-0.96 – 1.80	0.550

7.2. Appendix 2. Continued.

Predictors	**Estimates**	**CI**	**p**
ALLAGEOVER45_MEAN_LAG2	-0.58	-1.75 – 0.60	0.335
EST_ELIG_RATE_MEAN_LAG2	0.22	-5.53 – 5.96	0.941
ELIG_NEVER_PCT_MEAN_LAG2	0.01	-0.02 – 0.03	0.585
OCC_RATE_MEAN_LAG2	1.34	-2.38 – 5.07	0.481
MSG_MATCHQUALITY_MEAN_LAG2	-0.85	-2.49 – 0.79	0.311
MSG_AGE_MEAN_LAG2	-0.10	-0.14 – -0.05	**<0.001**
MSG_INCOME_MEAN_LAG2	0.00	-0.00 – 0.00	0.919
EST_ELIG_15_49_ACS_MEAN_LAG2	-1.52	-5.16 – 2.12	0.413
TRIPS_LAG2	0.05	-0.02 – 0.12	0.191
FTFNOCONTACT_LAG2	-0.00	-0.01 – 0.01	0.368
FTFCONTACT_LAG2	-0.01	-0.06 – 0.05	0.820
FTFAPPT_LAG2	-0.17	-0.29 – -0.05	**0.005**
MAINIW_LAG2	-0.12	-0.24 – -0.00	**0.050**
FTFMAINCONCERN_LAG2	-0.01	-0.20 – 0.17	0.884
FTFMAINNI_LAG2	0.16	-0.38 – 0.70	0.561
FTFMAINNS_LAG2	2.11	-9.37 – 13.60	0.718
FTFSCRNIW_LAG2	-0.12	-0.16 – -0.07	**<0.001**
FTFSCRNCONCERN_LAG2	-0.02	-0.15 – 0.11	0.797
FTFSCRNNI_LAG2	0.33	-0.20 – 0.87	0.224
FTFSCRNNS_LAG2	0.02	-0.05 – 0.10	0.537
FTF_MAINNS_INEL_LAG2	-1.19	-3.04 – 0.66	0.208
ACTIVE_LINES_LAG2	0.03	0.03 – 0.04	**<0.001**
TEL_ALL_LAG2	-0.00	-0.02 – 0.01	0.498
Random Effects			
σ^2	98.58		
$\tau_{00 \text{newIwerID}}$	32.31		
ICC	0.25		
$\text{N}_{\text{newIwerID}}$	187		
Observations	8843		
Marginal R^2 / Conditional R^2	0.038 / 0.275		

In this table, the variable names are used. See Appendix 1 (Subsection 7.1) for a description of each variable.

8. References

Abu-Nimeh, S., D. Nappa, X. Wang, and S. Nair. 2008. . "Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy." 2008 Third International Conference on Availability, Reliability and Security, Barcelona, Spain, 4–7 March 2008 IEEE. Available at: https://ieeexplore.ieee.org/abstract/document/4529459 (accessed May 2020).

Axinn, W., C. Link, and R. Groves. 2011. "Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior." *Demography* 48(3): 1–23. DOI: https://doi.org/10.1007/s13524-011-0044-1.

Barber, J.S., Y. Kusunoki, and H.H. Gatny. 2011. “Design and Implementation of an Online Weekly Survey to Study Unintended Pregnancies: Preliminary Results.” *Vienna Yearbook of Population Research* 9: 327–334. DOI: https://doi.org/10.1553/populationyearbook2011s327.

Biemer, P.P., de Leeuw, E.D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., Tucker, C., and West, B.T. (Eds.). 2017. *Total Survey Error in Practice*. Hoboken, New Jersey: Wiley.

Biemer, P.P., and D. Trewin. 1997. “A Review of Measurement Error Effects on the Analysis of Survey Data.” In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (pp. 601–632). New York: Wiley.

Burger, J., K. Perryck, and B. Schouten. 2017. “Robustness of Adaptive Survey Designs to Inaccuracy of Design Parameters.” *Journal of Official Statistics* 33(3): 687–708. DOI: https://doi.org/10.1515/jos-2017-0032.

Chipman, H.A., E.I. George, and R.E. McCulloch. 2010. “BART: Bayesian Additive Regression Trees.” *The Annals of Applied Statistics* 4(1): 266–298. DOI: https://doi.org/10.1214/09-AOAS285.

Dorie, V., H. Chipman, R. McCulloch, A. Dadgar, R.C. Team, G.U. Draheim, M. Bosmans, C. Tournayre, M. Petch, and R. de Lucena Valle. 2019. “dbarts: Discrete Bayesian Additive Regression Trees Sampler.” Available at: https://CRAN.R-project.org/package = dbarts (accessed May 2020).

Durrant, G.B., O. Maslovskaya, and W.F. Smith Peter. 2017. “Using Prior Wave Information and Paradata: Can They Help to Predict Response Outcomes and Call Sequence Length in a Longitudinal Study?” *Journal of Official Statistics* 33(3): 801–833. DOI: https://doi.org/10.1515/jos-2017-0037.

Finamore, J., S. Coffey, and B. Reist. 2013. “National Survey of College Graduates: A Practice-Based Investigation of Adaptive Design.” Annual AAPOR Conference, May 16–19, 2013. Boston, MA, U.S.A.

Green, D.P., and H.L. Kern. 2012. “Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees.” *Public Opinion Quarterly* 76(3): 491–511. DOI: https://doi.org/10.1093/poq/nfs036.

Groves, R.M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly* 70(5): 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., and S.G. Heeringa. 2006. “Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(3): 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Kern, C., T. Klausch, and F. Kreuter. 2019. “Tree-Based Machine Learning Methods for Survey Research.” *Survey Research Methods* 13(1): 73–93. DOI: https://doi.org/10.18148/srm/2019.v1i1.7395.

Kirgis, N., and J. Lepkowski. 2013. “Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006-2010 National Survey of Family Growth.” In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter: 121–144. Hoboken, NJ: Wiley.

Kleven, Ø., J. Fosen, B. Lagerstrøm, and L.-C. Zhang. 2010. . "The Use of R-Indicators in Responsive Survey Design–Some Norwegian Experiences." Q2010 Conference, Helsinki, 3–6 May 2010. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/risq/kleven-2010b.pdf (accessed May 2020)

Laflamme, F., and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." Proceedings of the European Conference on Quality in Official Statistics, Helsinki, Finland, Helsinki, Finland, 3–6 May, 2010. Available at: https://q2010.stat.fi/media/presentations/1_Responsive_design_paper_london_event1_revised.doc.

Lewis, T. 2017. "Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol." *Journal of Official Statistics* 33(3): 601–624. DOI: https://doi.org/10.1515/jos-2017-0029.

Luiten, A., and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1): 169–189. DOI: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Lundquist, P., and C.-E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29(4): 557–582. DOI: https://doi.org/10.2478/jos-2013-0040.

Lynn, p. 2016. "Targeted Appeals for Participation in Letters to Panel Survey Members." *Public Opinion Quarterly* 80(3): 771–782. DOI: https://doi.org/10.1093/poq/nfw024.

Mohl, C., and F. Laflamme. 2007. "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." Joint Statistical Meetings, Salt Lake City, UT, 29 July–2 August, 2007. Available at: http://www.asasrms.org/Proceedings/y2007/Files/JSM2007-000421.pdf (accessed May 2020).

Paiva, T., and J.P. Reiter. 2017. "Stop or Continue Data Collection: A Nonignorable Missing Data Approach for Continuous Variables." *Journal of Official Statistics* 33(3): 579–599. DOI: https://doi.org/10.1515/jos-2017-0028.

Peytchev, A., R.K. Baxter, and L.R. Carley-Baxter. 2009. "Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73(4): 785–806. DOI: https://doi.org/10.1093/poq/nfp037.

Peytchev, A., E. Peytcheva, and R.M. Groves. 2010. "Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences." *Public Opinion Quarterly* 74(2): 319–327. DOI: https://doi.org/10.1093/poq/nfq002.

Plewis, I., and N. Shlomo. 2017. "Using Response Propensity Models to Improve the Quality of Response Data in Longitudinal Studies." *Journal of Official Statistics* 33(3): 753–779. DOI: https://doi.org/10.1515/jos-2017-0035.

Rao, R.S., M.E. Glickman, and R.J. Glynn. 2008. "Stopping Rules for Surveys with Multiple Waves of Nonrespondent Follow-Up." *Statistics in Medicine* 27(12): 2196–2213. DOI: https://doi.org/10.1002/sim.3063.

Rosen, J.A., J. Murphy, A. Peytchev, T. Holder, J. Dever, D. Herget, and D. Pratt. 2014. "Prioritizing Low Propensity Sample Members in a Survey: Implications for Nonresponse Bias." *Survey Practice* 7(1). DOI: https://doi.org/10.1.1.686.6795.

Schonlau, M,. and M.P. Couper. 2016. "Semi-Automated Categorization of Open-Ended Questions." *Survey Research Methods* 10(2): 143–152. DOI: https://doi.org/10.18148/srm/2016.v10i2.6213.

Sparapani, R.A., B.R. Logan, R.E. McCulloch, and P.W. Laud. 2016. "Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART)." *Statistics in Medicine* 35(16): 2741–2753. https://doi.org/ DOI: 10.1002/sim.6893.

Tabuchi, T., F. Laflamme, O. Phillips, M. Karaganis, and A. Villeneuve. 2009. "Responsive Design for the Survey of Labour and Income Dynamics." Statistics Canada Symposium. October 27–30, 2009. Gatineau, Québec, Canada. Available at: http://oaresource.library.carleton.ca/wcl/2016/20160811/CS11-522-2009-eng.pdf#page=149.

Tan, Y.V., C.A. Flannagan, and M.R. Elliott. 2018. "Predicting Human-Driving Behavior to Help Driverless Vehicles Drive: Random Intercept Bayesian Additive Regression Trees." *Statistics and Its Interface* 11(4): 557–572. DOI: https://doi.org/10.4310/-SII.2018.v11.n4.a1.

Tourangeau, R., J. Michael Brick, S. Lohr, and J. Li. 2017. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(1): 203–223. DOI: https://doi.org/10.1111/rssa.12186.

Wagner, J. 2019. "Estimation of Survey Cost Parameters Using Paradata." *Survey Practice* 12(1): 1–10. DOI: https://doi.org/10.29115/SP-2018-0036

Wagner, J., and K. Olson. 2018. "An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys." *Journal of Official Statistics* 34(1): 211–237. DOI: https://doi.org/10.1515/jos-2018-0010.

Wagner, J., and T.E. Raghunathan. 2010. "A New Stopping Rule for Surveys." *Statistics in Medicine* 29(9): 1014–1024. DOI: https://doi.org/10.1002/sim.3834.

Wagner, J., B.T. West, H. Guyer, P. Burton, J. Kelley, M.P. Couper, and W.D. Mosher. 2017. "The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth." In *Total Survey Error in Practice*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West. New York. Wiley.

West, B.T., and A.G. Blom. 2017. "Explaining Interviewer Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5(2): 175–211. DOI: https://doi.org/10.1093/jssam/smw024.

West, B.T., J. Wagner, F. Hubbard, and H. Gu. 2015. "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth." *Journal of Survey Statistics and Methodology* 3(2): 240–264. DOI: https://doi.org/10.1093/jssam/smv004.

West, B.T., J. Wagner, S. Coffey, and M.R. Elliott. 2019. "The Elicitation of Prior Distributions for Bayesian Responsive Survey Design." *Historical Data Analysis versus Literature Review*. Available at: https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf.

Received August 2019
Revised February 2020
Accepted May 2020

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 933–935, http://dx.doi.org/10.2478/JOS-2020-0044

Book Review

Jennifer Edgar[1]

Paul C. Beatty, Debbie Collins, Lyn Kaye, Jose-Luis Padilla, Gordon B. Willis, and Amanda Wilmot. *Advances in Questionnaire Design, Development, Evaluation and Testing*. 2019, Wiley, ISBN: 978-1-119-26362-3, 816 pages.

Researchers have been thinking about, and researching the most effective approaches to do questionnaire design, development, evaluation and testing for a long time. Since well before the original International Conference on Questionnaire Design, Development, Evaluation and Testing (QDET) was held in 2002, the quest to find the most effective and efficient approaches to collect survey data has been the focus of researchers in academia, government agencies and private companies across the globe. In 2016, many of these researchers came together to share findings and innovations, and readers are fortunate that the information shared at that second conference (QDET2) was pulled together into this book to document the advances in the field, as well as the thoughts on questions yet to be answered by some of the most prominent players in the field.

Advances in Questionnaire Design, Development, Evaluation and Testing starts as one would hope, laying out the current state of affairs and highlighting issues most likely to face the field in the future. The first chapter was a perfect introduction for the book. Willis provided historical context while looking into the future to foreshadow several of the upcoming chapters. Dillman stayed at the big-picture level in the second chapter, connecting QDET2 back to the original QDET by identifying some of the key issues speakers tackled. Readers may be struck by the number of issues that remain relevant today, and are addressed explicitly in the volume.

Three additional introductory chapters lay out the current thinking around questionnaire design and evaluation, painting a rich picture of what we know after all the years of questionnaire design and evaluation research and pointing out the areas we all still need to work through. It's certainly difficult to argue with authors such as Willis, Tourangeau or Dillman based on their status in the field; they and their co-authors do an excellent job laying out the current state of affairs.

Throughout Part One of the book, there is a careful balancing of optimism with how far the field has progressed (e.g., Dillman no longer feeling as frustrated with the state of questionnaire evaluation as he was after the first QDET) with caution (e.g., Willis noting that there is still work to be done before we can speak to "what is a good question"). These overview chapters provide the audience with a useful perspective with which to view the subsequent chapters which tackle specific design, evaluation and testing topics.

[1] Bureau of Labor Statistics, Office of Survey Methods Research, 2 Massachusetts Ave., NE Washington D.C., U.S.A. Email: edgar.jennifer@bls.gov

In response to the initial chapters of the book, I'd like to challenge both Part I authors, and readers, to think about that question. If, after more than 20 years, evaluating the quality of a question, or a survey is elusive, what does that mean? Perhaps, the goal is not to identify a 'good' question per se, but instead to identify potential problems with the question, and then re-evaluate to determine if we reduced the likelihood of that problem occurring. Maybe we'll never get to a set of known 'good' questions, but instead can have a set of evaluation and testing techniques to get us to 'better.' Given the number of high quality studies looking at the impact of questionnaire design documented in this book, it is clear that researchers find design and evaluation efforts worthwhile, even if we're not necessarily able to quantify the improvement in question quality. While many of us would appreciate the opportunity to conduct research to fully evaluate our evaluation methods, in the current climate of declining response rates and other challenges, researchers are limited to addressing immediate problems like how to present a grid question on a mobile phone (Dale and Walsoe) or how to collect consistent information across cultures and languages (Smith). And, in the case of the authors represented in this volume, it is clear that they are tackling their real-world issues in a way that aims to address the overarching goal of administering high quality questionnaires.

The remaining chapters in this book provide readers information on a wide variety of subjects, and will undoubtedly serve as a valuable reference. Some chapters provide a useful overview of a topic, such as Yan and colleagues who provide a thorough overview of respondent burden including the relevant literature before adding some empirical evidence about which survey characteristics are predictors of burden. Others are more targeted, Nichols et al. who give readers a behind the curtain view of usability testing at the U.S. Census Bureau.

There are also articles that take a step back and provide insight into conducting and managing questionnaire evaluation research such as Stapelton et al. who talk about complex cognitive testing projects and Jans et al. who share an overview of how they leveraged multiple methods and iterative testing and even offer suggestions for future studies considering doing the same.

As a whole, the chapters cover a good mix of theory and research and provide readers a valued resource on a variety of topics. For readers who were not fortunate to attend QDET2, this volume provides a sense of the breadth and depth of questionnaire evaluation topics covered.

Reflecting on this book may leave the reader pondering "What issues will QDET3 tackle?" Will the Q no longer stand for questionnaire, instead representing the focus on evaluating Quality of blended data, administrative data, non-designed data, etc.? Or will we still be struggling with how to present questions to respondents to collect the most accurate information we can with minimal burden? While personally I am confident that surveys are here to stay, there are certainly opportunities to look at the techniques such as those presented in this book, that have been fine-tuned over the decades and identify ways to apply them to new types of information collections. As long as we seek to draw conclusions about people, regardless if it's by asking respondents directly or indirectly through non-designed data, we will always need to consider and evaluate the quality of information collected.

In summary, this book should be on the shelf of anyone who is actively working in the field. Not only does it provide survey practitioners with empirical information on a wide range of topics (e.g., measuring disability equality), it also provides thought-provoking information about methodology in the field (e.g., online pretesting methods or cross-cultural surveys). The combination of the two will serve readers well, regardless of their experience in the field, both novices and experts will find much to learn here. The topics and issues presented here, both those that seem resolved and the questions raised but not yet answered will undoubtedly stay relevant as we all continue to seek the most effective and efficient ways to design, develop, evaluate and test our questionnaires in the years to come.

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 937–940, http://dx.doi.org/10.2478/JOS-2020-0045

Book Review

Patricia Goerman[1]

Yuling Pan, Mandy Sha, and Hyunjoo Park. *The Sociolinguistics of Survey Translation*. 2020, New York: Routledge, ISBN 978-1-138-55087-2, 166 pages.

Over the last two decades, there have been increasing amounts of research on survey translation with the goals of: (1) including hard-to-count populations in surveys and (2) increasing comparability and quality of data collected via different language versions of the same survey instrument. In their new book, Pan, Sha and Park illustrate how the field of sociolinguistics can be a useful frame of reference for this type of work.

The book illustrates how to structure the whole survey translation process (design, translation, review and pretesting) in a sociolinguistic framework with the goal of improving survey translation quality. The authors discuss ways in which survey translation is different from other types of translation and propose ways in which the different practitioners involved in the process can work together using this framework. In a nutshell, the authors propose conducting translation at three different levels:

1. At the lexical level, in which the right words are chosen,
2. At the syntactic level in which appropriate use of grammar is the focus, and
3. At the pragmatic level where the social context and communicative effect of the words are considered.

Readers who would benefit from this book include the many actors involved in a survey translation process, such as survey sponsors, survey methodologists, translators, and subject matter experts. Professional translators and bilingual survey methodologists and students of either of these fields would also find this book useful. Finally, the book would serve as a great overview for a survey sponsor who wants to start the translation process of a given survey.

Those who work in the field of survey translation often encounter tensions related to the differing perspectives, goals, constraints and training across the various participants in the process. This book offers a useful conceptual framework to help navigate these issues. It also offers a frame of reference that could bridge the divides that sometimes occur between these groups to help them better work together.

The authors take a deep dive into the complexities involved in survey translation. A fundamental message that emerges is that merely sending a survey to a translator, receiving the translation back and placing the end product directly into the field is not

[1] U.S. Census Bureau, Center for Survey Measurement, 4600 Silver Hill Rd. 5K503, Washington, D.C., U. S. A. Email: patricia.l.goerman@census.gov

Acknowledgments: The views expressed in this review are those of the author and not necessarily those of the U.S. Census Bureau.

enough to ensure that the survey instrument will gather high quality, comparable data across languages. The book includes some technical language from the field of linguistics, and readers who have a basic understanding of those concepts might find the book more useful than others, though the authors do provide definitions for most of the terms they discuss. The book focuses on English as the base language prior to translation, being situated primarily in translation studies in the U.S.

The book is comprised of seven chapters and each one includes discussion questions that could be used in teaching a college or graduate level course. The first two chapters focus on an overview of the role of language in survey translation and an introduction to the field of sociolinguistics, which the authors define as "the study of language use in its social contexts" (page 1). The authors introduce key concepts and theories and provide a brief history of the disciplines of sociolinguistics and survey translation. They delve into the important topic of direct translation versus "adaptation," which can be defined as making modifications to the wording or structure of a question to better fit a different language or cultural group.

The third chapter focuses on questionnaire translation. The focus is on linguistic rules, social practices and cultural norms and the chapter offers a useful conceptualization of why survey translation can be difficult. The chapter talks about common pitfalls in translation, such as how to handle active versus passive voice structures and questions that include complex modifying clauses. The chapter also discusses issues of social practice that can trip up respondents even when accurate and technically correct translations are used. The authors discuss what to do when social practices are different across language and cultural groups. Finally, they talk about cultural norms such as politeness and coherence that may vary across cultures.

Chapter 4 moves into the idea of what the authors call "translation beyond words." Here they delve into issues such as how to handle languages that vary in writing right to left, the type of alphabet or writing system used and languages that have simplified versus traditional forms. This chapter looks at the issue of mode differences in surveys and points to a number of special topics for consideration when translating self-administered internet surveys.

The fifth chapter moves beyond the survey questionnaire itself and discusses the topic of translation of supplementary survey materials of two broad types: (1) survey letters, brochures and other respondent facing materials outside of the questionnaire itself and (2) study materials for use in pretesting translations with respondents. The first part of the chapter gives a nice overview of why "word for word" translation of survey letters often does not achieve the goal of convincing respondents to participate in the survey and provides examples of technically accurate translations that respondents do not interpret as intended due to social or cultural differences. In the end, the authors advocate the idea of a new approach to translation of survey letters, saying that they should be reconstructed using culturally appropriate discourse structures and letter writing styles rather than written and structured exactly like the source version.

The second part of Chapter 5 discusses the different types of research protocols and documents that often need to be translated in preparation for respondent pretesting using methods such as focus groups, cognitive interviews and usability testing. The authors provide a very useful discussion of three challenges that often arise:

1. The need to standardize in surveys,
2. Ensuring that surveys can generate the type of data intended, and
3. Cross-cultural differences in communication norms.

The book offers concrete types of probing questions to use and ways to test these concepts. However, the actual probe wording in other languages is not always included in the book, so bilingual survey methodologist readers who want to improve their methodology based on the authors' recommendations will have to seek the non-English probe wording elsewhere.

Chapter 6 moves past translation methods and focuses on ways to review and improve upon the initial draft version of translated survey instruments. It starts with a discussion of expert review and finishes with a discussion of respondent pretesting. The authors include an important discussion of why review by experts alone is not enough and why respondent testing is also needed. They recommend use of a comprehensive protocol for respondent pretesting but the book itself does not include sample protocols or say where to find examples. Novice readers who would like to implement this method will need to seek example protocols from elsewhere or hire researchers with background/expertise in the respondent translation pretesting field. The authors recommend making adjustments to research protocols based on cultural issues but the book itself does not contain enough information for someone to fully design such a protocol without use of outside resources.

The book includes several appendices. The first contains helpful examples of English language probing questions for use in respondent pretesting. The second contains instructions and a template for use in conducting expert reviews of a translation. A third appendix includes an English language example of an interactive practice session to help respondents understand the pretesting task. The last appendix contains a coding scheme that can be used to analyze issues uncovered in respondent pretesting.

One of the biggest contributions this book makes is providing the reader with an understanding of the reasons why much more work is required after the initial translation process to ensure functional equivalence across different language versions of a survey. It gives the reader an in-depth understanding of the steps to take and the basics of how to do this. However, it does not give enough information as a standalone for an individual translator, bilingual survey methodologist or survey sponsor who is new to this area to fully design and carry out a project of this sort. At the same time, the awareness the book brings could help with project planning, including helping people to plan for enough time and resources to allow for a high quality translation.

A practical question that remains unanswered is what type of practitioner could best implement and oversee a process of survey translation using a sociolinguistic framework in a typical survey organization. The authors state that most translators do not have training to do survey translation in this manner. There are presumably also a limited number of bilingual survey methodologists with a sociolinguistics background or sociolinguists with survey expertise. This type of book might encourage more students and agencies to move in the direction of developing this type of expertise. The book addresses the need for training of the translators and other participants in the process but it leaves unspoken who could best design and conduct such a training and when they should be brought into the process. Should researchers with this expertise run the whole translation process? Should they be brought in after an initial translation has already been done to oversee expert review and testing? Where can survey sponsors find this type of expertise? These are issues that could use further elaboration.

An additional question that the book mentions, though rather briefly, is the need to document changes made to translations based on adaptation from the original source version of materials. It is often the case that survey questions and materials are used repeatedly in different waves or iterations of a survey and that minor changes are made to the source wording between rounds. It can be difficult to ensure that adaptations made through review and testing are well documented and retained when materials need to be sent to a translator for updates. Detailed recommendations on how to best document and build on this type of work is an issue that remains to be explored.

The final chapter of the book reiterates the authors' useful framework to conceptualize the struggles that often arise during survey translation projects. The survey translation process can involve many practitioners with different backgrounds, who have competing priorities and constraints. Tensions between survey sponsors, translators, survey methodologists, subject matter experts and linguists can be difficult to resolve when each is approaching the process from a different perspective. The authors provide a nice overarching theory that could serve to bridge the gaps between these disparate groups.

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 941–943, http://dx.doi.org/10.2478/JOS-2020-0046

Book Review

Katherine Jenny Thompson[1]

Paul J. Lavrakes, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady West. *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*. 2019, Wiley, ISBN: 978-1-119-08374-0, 544 pages.

Qualitative and quantitative research are entrenched in survey methodology. Certainly, qualitative research tools that use small datasets such as focus groups and onsite visits of businesses provide invaluable information in assessing *construct validity* in survey instrument design, as does usability testing. However, these tools are limited, in the sense that they rely on observational data, are purposively gathered, and should not be used for any form of causal inference. At the opposite end of the spectrum, large amounts of data can be easily gathered via the internet using probability or nonprobability samples such as online convenience panels; the resulting statistics are time-dependent measures, generally considered to be externally *valid*. Again, however, these vehicles do not allow for cause-and-effect analysis.

Experimental Methods in Survey Research is an extensive collection of papers gathered together for one express purpose: to convince survey researchers to utilize experimental designs combining random assignment of subjects to treatment within random (probability) samples. *Internal validity* is attained via the random assignment of treatments, using blocking in the assignment process to control for "known" factors. Using a probability sample combined with successful recruitment methods should achieve *external validity*, assuming that the sampling frame covers the target population (i.e., has low *coverage error*) and that the response mechanism is *ignorable* (i.e., respondents are a random subsample). In assembling and organizing these chapters, the editors "embraced both the Campbell and Stanley validity framework" for research studies (statistical conclusion validity, construct validity, external validity, and internal validity) as well as the key components of the total survey error (TSE) framework, specifically coverage error, sampling error, nonresponse error, and measurement error. Thanks to the extensive usage of computer-assisted technologies in survey collection and the prevalence of internet access in the majority of target populations, the editors argue that there is little or no cost increase in embedding true experiments within ongoing programs, and therefore such tests should be carefully planned and executed on a wide scale in common practice.

The book is divided into nine topical areas, comprising twenty-three chapters (the first chapter is an introduction). Each chapter covers one experimental research topic, providing a literature review and presenting at least one case study. Not perhaps

[1] U.S. Census Bureau, Economic Statistical Methods Division, 4600 Silver Hill Road, Washington, D.C., 20233, U.S.A. Email: Katherine.j.thompson@census.gov.

Acknowledgments: Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

unexpectedly, the majority of the material covers issues related to household surveys; establishment surveys are limited to one chapter. Given this framework, there is a high percentage of real estate dedicated to attaining appropriate within-household coverage, and potential interviewer effects in terms of participation, response (unit and item), and experiment implementation. An equally high percentage of real estate is dedicated to assessing mode effects. In many of the presented studies, satisficing is a major quality concern, and response and completion rates are the primary measurement of treatment effects.

There are compelling arguments for embedding designed experiments into ongoing programs, especially probability surveys. However, there are other considerations that are highlighted in the book. Coverage of the broad population might not be an issue, but assessment of subdomain-specific effects might be. For example, comparisons between gender-specific or age-specific estimates could be made difficult by imbalanced samples. Interviewer effects can impact field experiments: is it better to assign one treatment per interviewer or all treatments to all interviewers? The scope of the experiment is important. Survey managers are unlikely to implement an experiment that might lower the overall response rate or noticeably affect the quality of the final estimates. For these practical reasons, embedded experiments might be confined to a portion of the survey sample such as small businesses or farms in an establishment survey. Carryover effects need to be considered with experiments embedded in longitudinal surveys; introducing or removing a treatment from a longitudinal panel could likewise have effects on response and/or quality and could lead to confounding. Telephone or internet collection could reduce or avoid interviewer effects. However, telephone-only and internet-only sample designs can be subject to frame coverage imbalances. Furthermore, response and completion rates for internet surveys tend to be very low. Each of these topics is discussed in detail in the context of experiments embedded in a variety of surveys.

As an overview of quantitative research applied to survey methodology, the book is a success. The variety of topics is comprehensive, and the literature overviews are generally very informative. Indeed, the authors provide thorough and understandable discussions of their study problem(s), the experimental designs, and the tested treatments. As a survey statistician – not a survey methodologist – I found the background sections useful; I imagine that they would be likewise invaluable to the novice survey methodologist. The inevitable compromises between implementing (statistically optimal) fully factorial designs and the practically feasible partially factorial designs are thoughtfully presented.

On the other hand, I had difficulty digesting other aspects of the material. First, low response rates and completion rates are endemic throughout the presented case studies, with response rates often well below 50%, generally closer to 25% with one (admittedly extreme) case study examining online panel data with a cumulative response rate of 6.1%. I found it difficult to justify the external validity of such experiments, without making heroic assumptions about nonresponse bias – that cannot be validated in practice. Many of the presented studies offer inconclusive results and attempt to limit their analyses, but there were chapters that presented large tables of contrast tests performed within the same experiment highlighting significant differences verging on p-hacking. The majority of studies accounted for complex designs in their analyses and made use of reweighting to correct for coverage; one chapter explicitly states that nonresponse effects in probability surveys are “ignorable or addressable through poststratification.” However, technical explanations were fairly limited, and most chapters provided very little concrete guidance in implementation for the novice survey methodologist or survey statistician and only one chapter providing formulae.

Ultimately, the book provides a snapshot of the *current* state of experimental design in probability samples. It is not a textbook, and it does not attempt to provide extensive historic context (for example, the numerous CATI/CAPI survey tests administered by the Census Bureau in the 1990s). In structure, it resembles a dedicated special issue in a survey methodology journal, with thoughtful introductions for each topic provided by the editors. It is well-organized and for the most part, well-written. The discussions of the tested hypotheses are explicitly laid out and grounded in theory. That said, there are severe drawbacks in presenting experimental results from studies plagued by low response rates, even when corrective weighting and appropriate statistical test modifications are employed, especially when analyzed from a total survey error perspective. While advocating for more experiments embedded in surveys, the book also illustrates the challenges in doing it well.

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 945–952, http://dx.doi.org/10.2478/JOS-2020-0047

Editorial Collaborators

The editors wish to thank the following referees and guest editors of theme issues who have generously given their time and skills to the Journal of Official Statistics during the period 1 October 2019 to 30 September 2020. An asterisk indicates that the referee served more than once during the period.

Abe, Naohito, Hitotsubashi Daigaku Institute of Economic Research, Tokyo, Japan
Abel, Guy, Asian Demographic Research Institute, Shanghai, China
Abraham, Katharine, Joint Program in Survey Methodology, College Park, Maryland, U.S.A.
Alaimo Di Loro, Pierfrancesco, La Sapienza – University of Rome, Rome, Italy
Ali, Sajid, Bocconi University, Milan, Italy
Andersson, Per Gösta, Stockholm University, Stockholm, Sweden*
Antoun, Christopher, University of Michigan, Ann Arbor, Michigan, U.S.A.
Argerich, Luis, University of Buenos Aires, Buenos Aires, Argentina
Arora, Sanjay, Ernst and Young LLP, Washington, D.C., U.S.A.
Ashmead, Robert, Ohio Colleges of Medicine Government Resource Center, Columbus, Ohio, U.S.A.
Axelson, Martin, Statistics Sweden, Örebro, Sweden
Bacchini, Fabio, Italian National Institute of Statistics, Rome, Italy*
Baffour, Bernard, University of Queensland, Brisbane, Australia*
Bakker, Bart, Statistics Netherlands, The Hague, the Netherlands
Balk, Bert, Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands*
Bashir, Shakila, Forman Christian Collage, Lahore, Punjab, Pakistan
Bates, Nancy, U.S. Census Bureau, Washington, D.C., U.S.A.
Bavdaž, Mojca, University of Ljubljana, Ljubljana, Slovenia
Beaumont, Jean-Francois, Statistics Canada, Ottawa, Canada
Benedetti, Roberto, University of Chieti Pescara, Pescara, Italy*
Beręsewicz, Maciej, Poznań University of Economics and Business, Wielkopolska, Poland*
Beresovsky, Vladislav, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.*
Berglund, Frode, Statistics Norway, Oslo, Norway
Bersimis, Sotiris, University of Piraeus, Piraeus, Greece
Bethlehem, Jelke, Leiden University, Leiden, the Netherlands*
Beyler, Amy, Mathematica Policy Research Health, Arlington, Virginia, U.S.A.
Białek, Jacek, University of Lodz, Lodz, Poland*
Biggeri, Luigi, University of Florence, Florence, Italy*
Bijlsma, Ineke, Maastricht University, Maastricht, the Netherlands*

Bivand, Roger, NHH Norwegian School of Economics, Bergen, Norway
Bohk-Ewald, Christina, Max Planck Institute for Demographic Research, Rostock, Germany
Bocquier, Philippe, Catholic University of Louvain, Louvain-la-Neuve, Belgium*
Boldsen, Carsten, UNECE, Geneva, Switzerland
Boonstra, Harm Jan, Statistics Netherlands, Heerlen, the Netherlands*
Bottone, Marco, Bank of Italy, Rome, Italy
Braaksma, Barteld, Statistics Netherlands, Utrecht, the Netherlands
Briceno-Rosas, Roberto, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany*
Brick, Michael, Westat, Rockville, Maryland, U.S.A.*
Breidt, Jay, Colorado State University, Colorado, U.S.A.
Brown, James University of Technology Sydney, Broadway, Australia
Brunåker, Fabian, Valueguard Index Sweden, Uppsala, Sweden*
Brunori, Paulo, University of Bari Aldo Moro, Bari, Italy
Bryant, John, Bayesian Demography Limited, Russley, Christchurch, New Zealand*
Buono, Dario, Eurostat, Mamer, Luxembourg
Burgard, Jan, University of Trier, Trier, Germany*
Cage, Robert, Bureau of Labor Statistics, Washington D.C., U.S.A.
Calviño, Aida, Universitat Rovira i Virgili, Madrid, Spain
Capecchi, Stefania, University of Naples Federico II, Naples, Italy*
Chen, Sixia, Westat, Rockville, Maryland, U.S.A.*
Cheng, Hao, National Academy of Innovation Strategy, Chaoyang, China
Coelho, Edviges, Statistics Portugal, Lisbon, Portugal
Cohen, Robin, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.
Coffey, Stephanie, U.S. Census Bureau, Washington, D.C., U.S.A.*
Conti, Pier Luigi, La Sapienza – University of Rome, Rome, Italy
Coquet, Francois, ENSAI, Bruz, France
Creel, Darryl, RTI International, Rockville, Maryland, U.S.A.*
Cruyff, Maarten, Utrecht University, Utrecht, the Netherlands
Czajka, John, Mathematica Policy Research, Washington. D.C., U.S.A.*
D'Alberto, Riccardo, University of Bologna, Bologna, Italy
Daas, Piet, Statistics Netherlands, Heerlen, the Netherlands
Dalla Chiara, Elena, University of Verona, Verona, Italy
Davidson, Russell, McGill University, Montreal, Quebec, Canada
Davern, Michael, NORC/University of Chicago, Chicago, Illinois, U.S.A.
De Coninck, David, Catholic University of Leuven, Leuven, Belgium*
De Haan, Jan, Statistics Netherlands, The Hague, the Netherlands*
De Leeuw, Edith, University of Utrecht, Utrecht, the Netherlands
Dennis, Michael, National Opinion Research Center, AmeriSpeak, Sunnyvale, California, U.S.A.
Deutsch, Tomi, Zavod Republike Slovenije za šolstvo, Ljubljana, Slovenia
De Waal, Ton, Statistics Netherlands, The Hague, the Netherlands
Di Cecco, Davide, La Sapienza – University of Rome, Rome, Italy*
Di Consiglio, Loredana, Italian National Institute of Statistics, Rome, Italy

Diewert, Erwin, University of British Colombia, Vancouver, British Columbia, Canada*
Di Fonzo, Tommaso, University of Padova, Padova, Italy*
Di Gennaro, Luca, National Statistics Office, Valletta, Malta*
Di Iorio, Francesca, University of Naples Federico II, Naples, Italy*
Di Zio, Marco, Italian National Institute of Statistics, Rome, Italy
Doidge, James, Intensive Care National Audit and Research Centre, London, UK
Drechsler, Jörg, Institute for Employment Research, Nuremberg, Germany
Dykema, Jennifer, University of Wisconsin, Madison, Wisconsin, U.S.A.*
Eck, Daniel, University of Illinois, Champaign, Illinois, U.S.A.
Eggleston, Jonathan, U.S. Census Bureau, Washington, D.C., U.S.A.
Elliott, Duncan, Office for National Statistics, Newport, UK
Erciulescu, Andreea, National Institute of Statistical Sciences, Washington, D.C., U.S.A.*
Evangelista, Rui, Eurostat, Luxembourg, Luxembourg
Evans, Thomas, U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.
Fabrizi, Enrico, Catholic University, Piacenza, Italy
Falorsi, Piero, Italian National Institute of Statistics, Rome, Italy
Farrugia, Naomi, National Statistics Office Malta, Valletta, Malta
Fischer, Mirjam, German Institute for Economic Research, Berlin, Germany
Fink, Paul, Ludwig Maximilian University of Munich, Munich, Germany
Flower, Tanya, Office for National Statistics, Newport, UK
Fonseca, Thais, University of Warwick, Coventry, UK
Fuller, Wayne, Iowa State University, Ames, Iowa, U.S.A.*
Gemenis, Kostas, Max Planck Institute for the Study of Societies, Cologne, Germany*
Geßendorfer, Jonathan, United Nations Statistics Division, New York, U.S.A.
Giesen, Deirdre, Statistics Netherlands, Heerlen, the Netherlands*
Gile, Krista, University of Massachusetts, Amherst, Massachusetts, U.S.A.
Goicoa, Tomás, Public University of Navarre, Navarre, Spain
Goldhammer, Bernhard, European Central Bank, Frankfurt am Main, Germany
Golinelli, Daniela, RAND Corporation, Santa Monica, California, U.S.A.
Graham, Patrick, Statistics New Zealand, Christchurch, New Zealand*
Grazzini, Jacopo, Eurostat, Luxembourg, Luxembourg
Greselin, Francesca, University of Milan-Bicocca, Milan, Italy
Groenitz, Heiko, Philipps-University Marburg, Marburg, Germany
Gweon, Hyukjun, University of Waterloo, Waterloo, Ontario, Canada
Hanif, Muhammad, National College of Business Administration and Economics, Lahore, Pakistan
Haraldsen, Gustav, Statistics Norway, Kongsvinger, Norway
Haslett, Stephen, Massey University, Palmerston North, Manawatu, New Zealand
He, Yulei, National Center for Health Statistics, CDC, Hyattsville, Maryland, U.S.A.*
Heckathorn, Douglas, Cornell University, Ithaca, New York, U.S.A.
Hedlin, Dan, Stockholm University, Stockholm, Sweden*
Heuchenne, Cedric, University of Liege, Liège, Belgium*
Heumann, Christian, Ludwig Maximilian University of Munich, Munich, Germany
Hill, Robert, University of Graz, Graz, Austria*
Himelein, Kristen, World Bank, Washington, D.C., U.S.A.

Honchar, Oksana, Australian Bureau of Statistics, Sydney, Australia
Hu, Jingchen, Vassar College, Poughkeepsie, New York, U.S.A.
Humer, Stefan, Vienna University of Economics, Vienna, Austria*
Iseh, Matthew, Akwa Ibom State University,Akwa Ibom, Nigeria*
Jansen, Ronald, UN Statistics Division, New York, U.S.A.*
Jaspers, Eva, Utrecht University, Utrecht, the Netherlands
Johansson, Anton, Statistics Sweden, Örebro, Sweden
Jones, Jacqui, Australian Bureau of Statistics, Belconnen, Australia
Joyce, Patrick, U.S. Census Bureau, Washington, D.C., U.S.A.
Joye, Dominique, University of Lausanne, Lausanne. Switzerland
Junker, Christoph, Federal Statistical Office Health, Neuchâtel, Switzerland
Karlberg, Forough, Luxembourg Statistical Services, Niederanven, Luxembourg*
Karr, Alan, RTI International, Durham, North Carolina, U.S.A.*
Karanka, Joni, Office for National Statistics, Cardiff, UK*
Kavee, Andrew, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.
Kennedy, Lauren, Monash University, Melbourne, Australia
Kern, Christoph, University of Mannheim, Mannheim, Germany*
Khan, M.G.M, University of the South Pacific, Suva, Fiji*
Kim, Jae-Kwang, Iowa State University, Ames, Iowa, U.S.A.*
Kinyon, David, Energy Information Agency, Washinton, D.C., U.S.A.
Kirby, Graham, University of St. Andrews, St. Andrews, Fife, UK
Klee, Mark, U.S. Census Bureau, Washington, D.C., U.S.A.
Kleven, Øyvin, Statistics Norway, Oslo, Norway
Koerner, Thomas, German Federal Statistical Office, Wiesbaden, Germany
Kolenikov, Stanislav, Abt SRBI, Silver Spring, Maryland, U.S.A.
Konjin, Paul, Eurostat, Luxembourg, Luxembourg
Kott, Phillip, RTI International, Derwood, Maryland, U.S.A.
Kowarik, Alexander, Statistics Austria, Vienna, Austria*
Koyuncu, Nursel, Hacettepe University, Ankara, Turkey
Kristoffersson, Ida, Swedish National Road and Transport Research Institute, Stockholm, Sweden
Lamboray, Claude, Eurostat, Luxembourg, Luxembourg
Larsen, Michael, George Washington University, Rockville, Maryland, U.S.A.*
Laud, Purushottam, Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A.*
Lee, Sunghee, University of Michigan, Ann Arbor, Michigan, U.S.A.
Lindholm, Mathias, Stockholm University, Stockholm, Sweden
Lineback, Fane, U.S. Census Bureau, Washington, D.C., U.S.A.
Lipps, Oliver, University of Lausanne, Lausanne, Switzerland*
Little, Roderick, University of Michigan, Ann Arbor, Michigan, U.S.A.*
Liu, Mingnan, SurveyMonkey, Palo Alto, California, U.S.A.*
Loong, Bronwyn, Australian National University Canberra, Australia
Loosveldt, Geert, Catholic University of Leuven, Leuven, Belgium*
Lundquist, Peter, Statistics Sweden, Solna, Sweden
Luiten, Annemieke, Statistics Netherlands, Heerlen, the Netherlands

MacDonald, Angus, Heriot-Watt University, Edinburgh, UK
MacFeely, Steve, UN Conference on Trade and Development, Geneva, Switzerland*
Magnusson, David, Valueguard Index Sweden, Uppsala, Sweden
Maitland, Aaron, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.
Malmros, Jens, Statistics Sweden, Solna, Sweden
Maltagliati, Mauro, University of Florence, Florence, Italy
Maples, Jerry, U.S. Census Bureau, Washington, D.C., U.S.A.*
Maslovskaya, Olga, University of Southampton, Southampton, UK
Massing, Natascha. GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany
Mazzuco, Stefano, Padua University, Padua, Italy
McElroy, Tucker, U.S. Census Bureau, Washington, D.C., U.S.A.*
Mecatti, Fulvia, University of Milan-Bicocca, Milan, Italy
Mercer, Andrew, Pew Research Center, Washington, D.C., U.S.A.
Misson, Sebastian, The Social Research Centre, Melbourne, Australia
Mitra, Robin, University of Southampton, Southampton, UK
Modugno, Lucia, Bank of Italy, Rome, Italy
Moradi, Abbas, Statistical Centre of Iran, Tehran, Iran
Morales Gonzalez, Domingo, Miguel Hernández University of Elche, Elche, Spain
Mothashami, Gholamreza, Ferowsi University of Mashhad, Iran*
Moultrie, Tom, University of Cape Town, Rondebosch, South Africa
Mukherjee, Diganta, Indian Statistical Institute, Kolkata, India*
Mukhopadhyay, Pushpal, SAS Institute Inc., Cary, North Carolina, U.S.A.
Mule, Vincent, U.S. Census Bureau, Suitland, Maryland, U.S.A.
Neri, Laura, University of Siena, Siena, Italy*
Neumann, Robert, Dresden University of Technology, Dresden, Germany
Nicholson, James, Durham University, Durham, UK
Nishimura, Raphael, Abt SRBI, Los Angeles, California, U.S.A.*
Norberg, Anders, Statistics Sweden, Solna, Sweden
Ograjensek, Irena, University of Ljubljana, Ljubljana, Slovenia*
Olteanu-Raimond, Ana-Maria, IGN, Saint-Mandé, France
Oncel Cekim, Hatice, Hacettepe University, Ankara, Turkey
Ogwang, Tomson, Brock University, St. Catharines, Ontario, Canada
Opsomer, Jean, Westat, Rockville, Maryland, U.S.A.
Oral, Evrim, Louisiana State University, New Orleans, Louisiana, U.S.A.
Orusild, Tiina, Statistics Sweden, Solna, Sweden
Osier, Guillaume, National Institute of Statistics and Economic Studies, Luxembourg, Luxembourg
Palm, Viveka, Statistics Sweden, Solna, Sweden
Pang, Osbert, U.S. Census Bureau, Washington, D.C., U.S.A.
Park, Mingue, Korea University, Seoul, Republic of Korea
Park, Minjeong, Statistical Research Institute, Seo-gu, Daejeon, Republic of Korea
Pedlow, Steven, NORC/University of Chicago, Chicago, Illinois, U.S.A.
Persson, Andreas, Statistics Sweden, Örebro, Sweden*
Pinheiro Jacob, Guilherme, Nossa Sra. das Graças, Manaus, Amazonas, Brazil
Planas, Christophe, Joint Research Centre of EC, Ispra, Varese, Italy*

Pascal, Joanne, U.S. Census Bureau, Washington, D.C., U.S.A.
Pratesi, Monica, University of Pisa, Pisa, Italy.
Presser, Stanley, University of Maryland, College Park, Maryland, U.S.A.
Proietti, Tommaso, University of Rome, Rome, Italy*
Psarakis, Stelios, University of Athens, Athens, Greece*
Quick, Harrison, Drexel University, Philadelphia, Pennsylvania, U.S.A.
Rambaldi, Alicia, University of Queensland, Brisbane, Australia
Rau, Roland, Max Planck Institute for Demographic Research, Rostock, Germany
Raymer, James, Australian National University, Canberra, Australia*
Righi, Paolo, Italian National Institute of Statistics, Rome, Italy*
Robbins, Michael, University of Missouri, Columbia, Missouri, U.S.A.
Roberson, Andrea, U.S. Census Bureau, Washington, D.C., U.S.A.*
Robison, Edwin, U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.
Rocco, Emilia, University of Florence, Florence, Italy
Rothe, Patrick, Bavarian State Office for Statistics, Fuerth, Germany*
Rothschild, David, Microsoft Research, New York, New York, U.S.A.
Russ, Daniel, National Institutes of Health, Bethesda, Maryland, U.S.A.
Salvati, Nicola, University of Pisa, Pisa, Italy
Scannapieco, Monica, Italian National Institute of Statistics, Rome, Italy
Scheffer, Fredrik, Statistics Sweden, Solna, Sweden*
Scherpenzeel, Annette, Technical University of Munich, Munich, Germany
Schliep, Erin, University of Missouri, Columbia, Missouri, U.S.A.
Schmid, Timo, Free University of Berlin, Berlin, Germany*
Schmidt, Tobias, Deutsche Bundesbank, Frankfurt am Main, Germany
Scholtus, Sander, Statistics Netherlands, The Hague, the Netherlands
Schonlau, Matthias, University of Waterloo, Waterloo, Ontario, Canada*
Schoumaker, Bruno, Catholic University of Louvain, Louvain-la-Neuve, Belgium
Schouten, Berry, Statistics Netherlands, The Hague, the Netherlands
Sebastiani, Fabrizio. Qatar Computing Research Institute, Doha, Qatar
Sengupta, Manisha, National Center for Health Statistics, CDC, Hyattsville, Maryland, U.S.A.*
Serfioti, Maria. European Commission (DG ESTAT), Luxembourg, Luxembourg
Seyb, Allyson, Statistics New Zealand, Christchurch, New Zealand
Shabbir, Javid, Quaid-i-Azam University, Islamabad, Pakistan*
Shin, Hee-Choon, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.*
Si, Yajuan, University of Michigan, Ann Arbor, Michigan, U.S.A.
Sief, Asghar, Bu-Ali Sina University, Hamedan, Iran
Silver, Mick, International Monetary Fund, Washington, D.C., U.S.A.
Singh, G.N., Indian Institute of Technology, Dhanbad, India
Singh, Sarjinder, Texas AM University-Kingsville, Kingsville, Texas, U.S.A.
Sixta, Jaroslav, University of Economics, Prague, Czech Republic
Smith, Duncan, University of Manchester, Manchester, UK*
Smith, Peter, University of Southampton, Southampton, UK
Sparapani, Rodney, Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A.*
Spoorenberg, Thomas, United Nations Population Division, New York, U.S.A.

Steinmetz, Stephanie, University of Lausanne, Lausanne, Switzerland
Steurer, Miriam, University of Graz, Graz, Austria
Ståhl, Olivia, Statistics Sweden, Solna, Sweden
Stinson, Martha, U.S. Census Bureau, Washington, D.C., U.S.A.
Stockhammar, Pär, Stockholm University, Stockholm, Sweden*
Stoop, Ineke, SCP Sociaal en Cultureel Planbureau, The Hague, the Netherlands
Struminskaya, Bella, Utrecht University, Utrecht, the Netherlands
Szymkowiak, Marcin, Poznan University of Economics and Business, Wielkopolska, Poland
Tanguy, Antoine, National Statistics Office, Valletta, Malta
Templ, Matthias, Statistics Austria, Vienna, Austria
Ten Bosch, Olav, Statistics Netherlands, The Hague, the Netherlands
Tent, Reinhard, Federal Statistical Office, Wiesbaden, Germany
Thalji, Lisa, RTI International, Research Triangle Park, North Carolina, U.S.A.
Thompson, Mary, University of Waterloo, Waterloo, Ontario, Canada
Tijdens, Kea, University of Amsterdam, Amsterdam, the Netherlands
Todorov, Valentin, UNIDO, Research and Statistics Vienna International Centre, Vienna, Austria
Tomadini, Andrew, Australian Bureau of Statistics, Belconnen, Australia
Tuoto, Tiziana, Italian National Institute of Statistics, Rome, Italy*
Umbach, Gaby, European University Institute, Florence, Italy*
Valliant, Richard, University of Michigan, Chevy Chase, Maryland, U.S.A.
Van den Brakel, Jan, Maastricht University School of Business and Economics, Maastricht, the Netherlands
Van der Laan, Jan, Statistics Netherlands, The Hague, the Netherlands
Van Loon, Ken, Statistics Belgium, Brussels, Belgium
Verma, Med Ram, Indian Veterinary Research Institute, Izatnagar, Bareilly, Uttar Pradesh, India*
Verzosa, Nina, University of Sydney, Sydney, Australia
Von Auer, Ludwig, University of Trier, Trier, Germany
Vozar, Ondrej, Czech Statistical Office, Prague, Czech Republic
Waldmann, Patrik, Swedish University of Agricultural Sciences, Uppsala, Sweden*
Walter, Paul, Free University of Berlin, Berlin, Germany
Wang, Yan, RMIT University, Melbourne, Australia*
Warren, John, University of Minnesota, Minneapolis, Minnesota, U.S.A.
Watson, Nichole, University of Melbourne, Melbourne, Australia
Wesolowski, Jacek, Central Statistical Office, Warszawa, Poland
Willenborg, Leon, Statistics Netherlands, The Hague, the Netherlands
Wiśniowski, Arkadiusz, University of Southampton, Southampton, UK
Wright, Tommy, U.S. Census Bureau, Washington, D.C., U.S.A.
Wu, Changbao, University of Waterloo, Waterloo, Ontario, Canada
Yildiz, Dilek, JOS-2019-0124, University of Southampton, Southampton, UK
Zabala, Felipa, Statistics New Zealand, Wellington, New Zealand
Zammit, Silvan, National Statistics Office, Valletta, Malta
Zenga, Michele, University of Milan-Bicocca, Milan, Italy

Zhang, Junni, Peking University, Beijing, China
Zhang Mark (Xichuan), Australian Bureau of Statistics, Belconnen, Australia
Zhang, Yunxi, University of Mississippi; Jackson, Mississippi, U.S.A.*
Zimmermann, Thomas, Federal Statistical Office, Wiesbaden, Germany

Journal of Official Statistics, Vol. 36, No. 4, 2020, pp. 953–956, http://dx.doi.org/10.2478/jos-2020-0048

Index to Volume 36, 2020

Contents of Volume 36, Numbers 1–4

Author Index

Book Reviews