Journal of Official Statistics
vol. 35, 1 (marzo 2019)

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 1–8, http://dx.doi.org/10.2478/JOS-2019-0001

Letter to the Editor

Extracting Statistical Offices from Policy-Making Bodies to Buttress Official Statistical Production

The importance of official statistics is increasing not only for effective and rational government operations and policies – its original use – but also for the efficient functioning of domestic and international markets, international collaboration and cooperation, scientific and technological progress, and for the functioning of the democratic system. Official statistics is a public good (Georgiou 2017) whose reliability and overall quality should be safeguarded and buttressed.

In a number of countries, statistical offices or bureaus that are entrusted with the production of the official statistics of the country, are part of policy-making institutions. Thus, as the United Nations Statistics Division notes on the basis of its latest global survey on the implementation of UN Fundamental Principles of Statistics (UNFP): "Some national statistical offices have a high degree of administrative independence, others are actually part of a ministry" (UNSD 2013). It should give us pause that the work and performance of these policy institutions and the politicians heading them are assessed, to a large extent, on the basis of the statistics produced by the statistical offices embedded in these institutions. This institutional setup increases risks to the implementation of statistical principles, including the principles of professional independence, impartiality and objectivity, and statistical confidentiality. Dependencies and conflicts of interest are inherent in this setup. Creating an administrative distance between policy-making institutions and statistical producers, by extracting statistical offices from policy-making bodies, is one of the necessary means of buttressing the professional independence and other critical aspects of the quality of official statistics for the long run. This is in the best interest of official statistical producers, but also – most importantly – in the best interest of the very wide variety of users of official statistics in modern society, including policy makers and political leaders.

1. Increase in Risks when Statistical Offices are Part of Policy-Making Bodies

The system of statistical offices or bureaus being part of policy-making bodies implies significant risks for the implementation of international statistical principles during the production of official statistics by such offices. There are risks to the statistical principles, including professional independence, impartiality and objectivity, and statistical confidentiality.

Risks are of different types and originate from various situations. There are two broad types of risks we are concerned with here: "pressure risk" and "political attack risk".

(i) "Pressure risk" is the risk that pressures will occur to circumvent statistical principles, either from outside the statistical perimeter (i.e., from persons/entities

that are users of official statistics such as policy makers, legislators, politicians, civil servants/administrators, market participants, academic researchers and the general public, or from upstream data providers (Georgiou 2018) or in the form of self-censorship.

a. "External pressure risk" is the risk of pressure from persons/entities outside the statistical perimeter on official statisticians to make decisions on the basis of nonstatistical considerations,
b. "Self-censorship risk" is the risk of pressure official statisticians may feel to engage in self-censorship and modify their statistical behavior without having received overt external pressure. They engage in self-censorship anticipating the sensibilities (perceived or actual) of policy makers or of others outside the statistical perimeter and, thus, allow nonstatistical considerations to affect their statistical decisions.

(ii) "Political attack risk" is the risk that official statisticians will be attacked by persons/entities in the political environment of official statistics production. The attack is usually justified by those that carry it out on the basis of allegations that official statisticians have succumbed to external pressure risk and self-censorship risk, or more generally that statisticians did not produce reliable and high quality statistics with independence, impartiality and objectivity.

All other things being equal, the above risks increase when statistical offices or bureaus are part of policy-making bodies and the official statistics producers report to the policy-making hierarchy.

The general argument is that a basic condition of existence for the long term robustness and sustainability of professional independence and of other fundamental statistical principles is the institutional independence of official statistics production (Georgiou 2018). Institutional independence is, by definition, incompatible with statistical offices and bureaus being part of policy-making bodies.

It should be noted that there is a fundamental distinction between the concept of institutional independence of official statistics production and the concept of professional independence of official statisticians.

Professional independence is when official statisticians (i) have the sole responsibility for deciding on statistical methods, standards and procedures, and on the content and timing of statistical releases; (ii) have responsibility for ensuring that statistics are developed, produced and disseminated in an independent manner; (iii) are free from political and other external interference in developing, producing and disseminating statistics; and (iv) carry out their compilation of statistics solely based on statistical principles and statistical legislation in force, without letting any other concerns sway their statistical decisions and without fear or favor in making their decisions. The definition of professional independence of official statistics offered here is informed by formulations of professionally independent behavior found in the European Statistics Code of Practice (Eurostat 2011) and in the ISI Declaration on Professional Ethics (International Statistical Institute 2010). **Institutional independence** of official statistics production is when the latter is independent from the executive, legislative or judicial branches of government.

The distinction between institutional independence and professional independence in official statistics is akin to the distinction between institutional independence and decisional independence in the case of the judiciary (Georgiou 2018). The judiciary's institutional independence (Lord Phillips 2011) is widely perceived to be a fundamental condition for its decisional independence. Similarly, the institutional independence of official statistics production should be seen as a fundamental condition of professional independence in official statistics production.

The specific argument underlying the thesis presented here is that the above noted "pressure risk" and "political attack risk" are mediated and amplified by hierarchical, administrative and resource dependencies of official statistics production on the policy-making body (and more broadly the executive branch of government) that the statistical office is part of. Some of the aspects of such dependencies are listed below:

- Hierarchical/authority/accountability relationships of officials in the statistical office with officials in the policy-making body,
- Conflation or amalgamation of any of the individual administrative and budgetary functions of the statistical office with those of the policy-making body,
- Control by the policy-making body and its policy (nonstatistical) officials of human resource issues (e.g., staff hiring, promotion, remuneration, terms and conditions of work), financial resource issues (e.g., access to approved budget funds, distribution of approved budget to expenditure lines, making expenditure commitments, financial administration, auditing and settling expenditures) and other resource issues (e.g., access to foreign aid, provision of information technology and related security) of the statistical office,
- Physical proximity/cohabitation of the statistical office with the policy-making body,
- Control of the selection, appointment, reappointment, remuneration and termination of the incumbency of the head of the statistical office by the policy-making body,
- Officials of the statistical office carrying out nonstatistical work/functions/tasks of the policy-making body they are part of along with their statistical ones,
- Assignment of parts of the statistical operations of the statistical office to the nonstatistical parts of policy-making institutions.

A number of these dependencies create or amplify conflicts of interest. The troubling role that conflicts of interest play in creating deep ethical dilemmas for various professions (e.g., accountants) has been recognized in behavioral economics (Ariely 2010).

2. Costs and Benefits of Statistical Offices as Part of Policy-Making Bodies

It is often argued that there are significant benefits of statistical offices being part of policy-making institutions. To make a rational choice on whether to extract statistical offices from policy-making bodies or leave them as they are, one would have to consider the costs and benefits of the two alternatives (Georgiou 2018).

The costs to statistics being part of a policy-making body are the costs of (i) the above listed risks materializing, (ii) the perception that the risks exist, and (iii) mitigating and managing the real and perceived risks. The economic and social, as well as political, costs that arise can be very large because the resultant official statistics actually impede, or are

perceived by a material share of the (domestic and international) public to impede, one or more of the following:

- the operation of the democratic system,
- the rationality and effectiveness of policy-making,
- international cooperation and the production of global public goods,
- the markets in operating effectively, adjusting in an orderly manner and leading to welfare maximization,
- scientific research and progress.

In addition, the costs to statistical offices (and the policy-making bodies to which they belong) of managing and mitigating the risks and the perception of these risks, all other things being equal, would be higher than when the statistical office is independent. The costs of effective supervision/checking of official statistics would also be higher. Furthermore, the costs (to the economy/society) of developing sources of information as an alternative to the official statistics (to address the above noted risks) would also tend to be higher than in a system where official statistics production takes place outside policy-making bodies.

Arguments for the benefits of statistical offices being part of policy-making institutions could include the following:

1. There is "access" to policy/decision makers that the head of the statistical office gets by being part of the hierarchy of a policy-making body. This is supposed to help the views of official statistics production be heard and serve the interests of statistics production (e.g., by protecting statistics production from various adverse legislative, budgetary and other policy developments).
2. There is greater access to other parts of the policy-making body, its human resources and its administrative data sources. Access to various parts of the civil service in the policy-making body is thought to help protect official statistics production from the various adverse developments mentioned above, as well as provide for human capital support in areas within statistics production.
3. Close relations with various levels of the administration of the policy-making body are also thought to facilitate administrative processing of various kinds of requests of statistics producers.
4. Access to administrative data sources is an important part of modern official statistics production and administrative proximity is supposed to facilitate such access.
5. Statistics production being part of a policy-making body that is a major user of these statistics is seen as a necessary condition for producing statistics relevant to the work of government.

To the above propositions regarding the benefits of statistical offices being part of policy-making institutions, one may juxtapose the following:

1. The head of the statistical office does not need to be part of the hierarchy of a policy-making institution to have access to policy/decision makers. Such access can be possible and can take place in an appropriate manner by providing for it in the law. Access is more likely to take place at the appropriate level of propriety and respect

for statistical independence when the statistical interlocutor is institutionally independent than when she/he is a subordinate and "reports" in the hierarchy of the policy-making institution that the statistics office is part of and is thus subject to clear conflicts of interest.

2. Accommodation of the appropriate interests and needs of official statistics production and protection from adverse legislative, budgetary and other policy developments would be best served with little risk of "quid pro quo", if it was provided for in law (for example, by providing in law for the role of statistics producers in the preparation of laws, including budgetary appropriations, with implications for statistical production) and the statistical office was not part of a policy-making institution. The closeness and collegiality of civil servants within the policy-making institution does not offer protection to statistical production from adverse legal/budgetary developments and administrative friction without also increasing – through, for instance, real and perceived conflicts of interest – "pressure risk" and "political attack risk", which undermine independence and other statistical principles.
3. Appropriate access to expertise and information existing in the policy-making institution does not have to go hand in hand with statistics being part of that institution. In any event, provision of expertise and information by a policy-making body is more likely to increase the risk of the policy-making/administrative perspective contaminating (again via conflicts of interest) the statistical approach when the statistical office is part of a policy-making institution. Official statistics production should and could have its own expertise in areas where it traditionally needs it and not be dependent on expertise existing in policy-making bodies.
4. Effective access to administrative data sources does not have to be mediated by the statistical office being part of a policy-making body. Access to administrative data sources is best achieved and, actually, statistical confidentiality best preserved when access to such data sources is provided for in law and the statistical office is not part of a policy-making body.
5. It is not necessary for official statistics production to be part of policy-making bodies in order to have a very attentive and responsive attitude by official statisticians towards the statistical needs of these policy-making bodies. Policy makers and their administrations do not need to have immediate physical and institutional access to official statisticians in order for the latter to be fully aware of and attentive to these important users' needs; the proper catering for such needs through appropriate arrangements (e.g., advisory user committees, user conferences, specialized user groups, and periodic, as well as ad hoc consultations with users) can be provided for in law. The risks are greater that user requests for "what" statistics are produced will get mixed up with conversations about "how" the statistics should be produced and "what outcome" the statistics should record when the statistical office is part of a policy-making body rather than outside it.

An argument that the cost of statistical offices being part of policy-making bodies is minimal is that the professional independence of official statistics production, even in this institutional setup, is secured through safeguards. Such safeguards may include: (i) provisions in law for the implementation of statistical principles in the national statistical

system; (ii) publicized policy commitments of governments to support confidence in official statistics; (iii) national institutions with the mandate to report on the implementation of statistical principles; (iv) provisions in law for the selection, fixed term, and termination of the incumbency of the head of the statistical office; (v) quality assurance of statistical output by supranational entities; and (vi) review of the implementation of statistical principles by supranational entities and processes.

First, it should be noted that such a set of safeguards is actually far from being in place in all countries and for all official statistics producers in national statistical systems. Second, any safeguards actually in place are often not in an appropriate, strong and effective form. This much can be gleaned even from the fragmentary information provided in the survey of the UNSD regarding the implementation of the UNFP (UNSD 2013). Moreover, a number of these safeguards are still, by and large, confined to and enabled in uncommonly strong regional partnerships of national statistical systems, such as in the European Union. Finally, there is evidence that very serious problems in the production of official statistics have occurred, even when such safeguards have been in place in some form. Greece's official statistics production leading up to the statistical crisis of 2009 (European Commission 2010) is just one example demonstrating the problems of the effectiveness of safeguards, even when such safeguards are actually in place.

Very importantly, while the above noted safeguards can help reduce the risks, the risks are not reduced to the degree that they would be if statistical offices were not part of policy-making institutions.

Another argument (in the case of certain countries or institutions) that the cost of statistical offices being part of policy-making bodies is minimal is that there has been a benign and benevolent environment in which official statistics production has taken place and risks such as "pressure risk" and "political attack risk" do not and cannot materialize.

Surely, having a benign environment and a benevolent approach towards official statistics by policy makers and politicians across the spectrum is desirable and gratifying when it happens. However, it is not the solid foundation on which to build statistical independence for the long run. In some way, the argument of the previous paragraph is akin to the argument that "there is no need for a judiciary separate from the sovereign because this king has traditionally been a benevolent and fair king in administering justice." In addition, history shows that even in countries with well-developed institutions (checks and balances) and a generally good statistical culture inside and outside the statistical perimeter, challenges to official statistics production and problems with adherence to statistical principles have occurred from time to time (Seltzer 1994), and there is no reason to believe that at some point they will not re-emerge. Thus a society needs to be ready for these moments and has to take steps to decrease the probability that the challenges and problems (i) will arise in any given period; and (ii) will be severe when they inevitably arise. These steps must include putting in place a proper and robust **institutional environment/basis** of official statistics production.

"Preparing for the worst, while hoping for the best" should be the general principle behind all choices regarding the appropriate institutional setting for official statistics production. In accordance with this principle, the specific institutional issue of statistical offices as part of policy-making institutions should be decided upon with a view to

preparing for all eventualities, even if the environment appears to be benign and has been benign for a while, because the environment can change and do so rapidly.

Official statistics production as part of policy-making bodies is a "legacy" institutional setup, with many risks and costs, and at best ambivalent benefits. What's more, the idea that statistical offices should be part of policy-making bodies is an anachronism; it belongs to another era, as does the idea that the exercise of judicial powers can be appropriately and sustainably carried out by the sovereign himself or by a judge in the court of the sovereign.

In conclusion, in a comparison of costs and benefits, we believe that the costs of statistical offices being part of policy-making bodies outweigh any benefits, and it is more effective and appropriate to extract statistical offices from policy-making bodies. However, extracting statistical offices from policy-making institutions does not mean that risks or a perception of risks will disappear, but only that they would materialize with a lower probability. Additional steps would need to be taken to decrease this probability further — full institutional independence of official statistics production along with other safeguards. Such steps would minimize dependencies and conflicts of interest and mitigate to the largest extent the effects of any remaining dependencies and conflicts of interest. This, in turn, would minimize the probability of risks materializing for statistics production. Meanwhile, extracting statistical offices from policy-making institutions is one first necessary step that has to be taken on the road to buttressing the long-term robustness and sustainability of professional independence and other fundamental statistical principles.

3. References

Ariely, D. 2010. "Gray areas in accounting." The Blog. Website, Retrieved from web May 1, 2018. Available at: http://danariely.com/2010/11/08/gray-areas-in-accounting/ (accessed January 2019).

European Commission. 2010. *Report on the Greek Government Deficit and Debt Statistics,* January. Available at: http://ec.europa.eu/eurostat/documents/4187653/6404656/COM_2010_report_greek/c8523cfa-d3c1-4954-8ea1-64bb11e59b3a (accessed January 2019).

Eurostat. 2011. Website. *European Statistics Code of Practice*. Available at: https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7 (accessed January 2019).

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). Available at: https://goo.gl/R8NKzU (accessed January 2019).

Georgiou, A. 2018. "Official Statistics Production Should be a Separate Branch of Government." *Statistical Journal of the IAOS* 34(2). Available at: https://content.iospress.com/download/statistical-journal-of-the-iaos/sji170399?id=statistical-journal-of-the-iaos%2Fsji170399 (accessed January 2019).

ISI, International Statistical Institute. 2010. Website. *International Statistical Institute Declaration on Professional Ethics*. Available at: https://www.isi-web.org/index.php/news-from-isi/34-professional-ethics/296-declarationprofessionalethics-2010uk?showall=1 (accessed January 2019).

Lord Phillips of Worth Matravers. 2011. *Judicial Independence and Accountability: A View from the Supreme Court*. Judicial Independence Research Project Launch. Available at: https://www.ucl.ac.uk/constitution-unit/events/judicial-independence-events/lord-phillips-transcript.pdf (accessed January 2019).

Seltzer, W. 1994. *Politics and Statistics: Independence, Dependence or Interaction?* United Nations, Department of Economic and Social Information and Policy Analysis, Working paper series No. 6. Available at: https://unstats.un.org/unsd/statcom/FP-Seltzer.pdf (accessed January 2019).

UNSD, United Nations Statistics Division. 2013. *Implementation of the Fundamental Principles of Official Statistics*, Background document, 44th Session of the Statistical Commission, 26 February – 1 March 2013. Available at: https://unstats.un.org/unsd/statcom/44th-session/documents/doc13/BG-FP-E.pdf (accessed January 2019).

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

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 9–30, http://dx.doi.org/10.2478/JOS-2019-0002

Consistent Multivariate Seasonal Adjustment for Gross Domestic Product and its Breakdown in Expenditures

Reinier Bikker[1], *Jan van den Brakel*[1], *Sabine Krieg*[1], *Pim Ouwehand*[1], *and Ronald van der Stegen*[1]

Seasonally adjusted series of Gross Domestic Product (GDP) and its breakdown in underlying categories or domains are generally not consistent with each other. Statistical differences between the total GDP and the sum of the underlying domains arise for two reasons. If series are expressed in constant prices, differences arise due to the process of chain linking. These differences increase if, in addition, a univariate seasonal adjustment, with for instance X-13ARIMA-SEATS, is applied to each series separately. In this article, we propose to model the series for total GDP and its breakdown in underlying domains in a multivariate structural time series model, with the restriction that the sum over the different time series components for the domains are equal to the corresponding values for the total GDP. In the proposed procedure, this approach is applied as a pretreatment to remove outliers, level shifts, seasonal breaks and calendar effects, while obeying the aforementioned consistency restrictions. Subsequently, X-13ARIMA-SEATS is used for seasonal adjustment. This reduces inconsistencies remarkably. Remaining inconsistencies due to seasonal adjustment are removed with a benchmarking procedure.

Key words: Seasonal adjustment; discrepancies; Kalman filter; multivariate structural time series models; X-13ARIMA-SEATS; benchmarking.

1. Introduction

Most national statistical institutes (NSIs) publish time series at an aggregated level and breakdowns in $K \geq 2$ domains, for instance the Gross Domestic Product (GDP) divided over industries or over expenditures.

It is common practice to adjust for seasonal and calendar effects. The latter are variations in time series that can be explained from variations in the calendar, such as working day patterns and national holidays. The aim of these adjustments is to make different reporting periods comparable. Seasonal and calendar adjustment procedures are generally based on univariate methods applied to the series of each publication domain separately. A consequence of such approaches is that adjusted figures at the aggregated level are not consistent with the sum of the adjusted figures of the underlying breakdown in K publication domains. This is a well-known problem and the status quo is that no

[1] Statistics Netherlands, P.O. Box 4481, 6401CZ Heerlen, The Netherlands. Emails: r.bikker@cbs.nl, ja.vandenbrakel@cbs.nl, s.krieg@cbs.nl, p.ouwehand@cbs.nl, and rhm.vanderstegen@cbs.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors are grateful to the unknown referees, the associate editor, Harm Jan Boonstra (Statistics Netherlands) and Jan van Dalen (Statistics Netherlands) for reading and commenting on a former draft of this article.

adequate solution exists. Eurostat's ESS guidelines on seasonal adjustment (Eurostat 2015) suggest computing the adjusted series at the aggregated level as a sum of the adjusted underlying domains, which is often referred to as the indirect approach. A drawback of this approach is that the most reliable estimates at the aggregated level are disregarded. Alternatively, if the discrepancies are small enough, they can be distributed by means of multivariate benchmarking techniques. These can be two-step procedures of benchmarking and reconciliation (Quenneville and Fortier 2006), or simultaneous methods as described in Di Fonzo and Marini (2011). In the Netherlands, however, the quarter-to-quarter changes in the discrepancies between the directly adjusted GDP and the sum of the adjusted series of its expenditures are often larger than the growth rate of GDP itself. This fact alone renders both remedies suggested by Eurostat unsuitable. For instance, in the first quarter after the 2013 recession, GDP grew by 0.1% if calculated directly, and by -0.9% if calculated indirectly.

The purpose of this article is to develop a method that attempts to make consistent seasonal adjusted series using a multivariate structural time series modelling approach. We focus on GDP and a breakdown in different expenditures. However, the proposed method is general and can be applied in any situation where consistent seasonal and calendar adjustment is required.

Another discrepancy is introduced by chain linking (see Bloem et al. 2001), in which GDP and its expenditures are calculated as chain volumes. In chain linking, series of a constant price level are constructed by "chaining" volume growth rates. These volume growth rates are calculated by dividing the nominal growth rate by a price factor. As each of the series in the breakdown of GDP has its own price factor, discrepancies arise between the sum of the expenditures and GDP itself.

In the Dutch case, the discrepancies from chain linking are typically smaller than the discrepancies introduced by the adjustments for seasonal and calendar effects. Moreover, we noted that the size of the discrepancies due to seasonal and calendar adjustment grew larger in the period 2009–2013. This period is characterized by rapid changes in seasonal patterns following the financial crisis in 2008/2009. The increasing size of the discrepancies eventually lead to complaints from users. The discrepancies were noted in the press, and professional users also asked how to interpret our published results.

Besides discrepancies, there are more quality aspects related to seasonal adjustment. When new data points become available and are added to the series, better estimates of the trend, the seasonal and calendar effect of all previous quarters, can be made. Therefore, revisions are inherent to seasonal and calendar adjustment, which are acceptable, as long as their size is not excessive.

Traditionally, the quality of seasonal and calendar adjustment is assessed using a well-defined set of criteria. In the case of X-13ARIMA-SEATS (U.S. Census Bureau 2015), the method used at Statistics Netherlands, these are the Q- and M-diagnostics. They are numerical scores given to properties, such as the amount of seasonality compared to noise and the rate at which the seasonal component changes over time. These criteria are optimised for each time series individually. After performing seasonal and calendar adjustment, the resulting discrepancies are calculated and only when these are very large, the seasonal and calendar adjustment may be changed. Revisions are monitored, but

never lead to changes in the setup of seasonal and calendar adjustment. So, the quality criteria that Statistics Netherlands traditionally applies, are (in order of decreasing importance):

1. Optimal quality diagnostics (specifically X-13ARIMA-SEATS's Q- and M-values).
2. Minimal statistical discrepancies between GDP and the sum of expenditures.
3. Minimal revision of the seasonal effect after adding new data points.

This is under the assumption that all criteria are within acceptable boundaries. As this was not the case in the Netherlands after the crisis in 2008/2009, the primary objective of the current research is to reduce the statistical discrepancy. This is achieved by introducing a multivariate approach. The consequence of the shift from an optimal univariate solution to a multivariate solution is that the seasonal and calendar adjustment of one series is influenced by another. Therefore, some aspects of the multivariate adjustment can be less optimal, when compared to the univariate case. However, a slightly lower quality (according to the Q- and M-diagnostics) can be equally acceptable for the users, as long as no residual seasonal effect can be found in a corrected series. Therefore, our goal is that, on average, the revisions and quality diagnostics should not deteriorate.

In this article, we describe two alternative approaches to adjustment for seasonal and calendar effects. The first approach applies a multivariate structural time series model to an aggregated series and its breakdown in K subseries. The model estimates all components subject to the constraint that the sum of the subseries is equal to the components of the aggregated series. Unfortunately, the results of this approach are not satisfactory due to numerical problems. Furthermore, the estimates for the seasonal components are considered to be too volatile. Therefore, a second approach is developed, which is based on a combination of a multivariate structural time series model and routines of X-13ARIMA-SEATS. Under this approach, the discrepancies are sufficiently reduced, while the size of the revisions is in the same order as before.

In Section 2 we will first define the problem in a more precise way. In Section 3, we present the multivariate state-space method for consistent seasonal adjustment. Section 4 discusses the results and finally, the article closes with a conclusion in Section 5.

2. Problem Definition

Statistics Netherlands publishes quarterly figures for GDP with both the final expenditures and the value added by industry as domains. These breakdowns are called the expenditure approach and the production approach. Both breakdowns are computed in constant prices (chain linked volumes) and in current prices. In this article, we will focus on the expenditure approach in constant prices.

In this article the aggregate $B1G$, the GDP, is itemized in the subseries $P7$ (imports), $P3_{S1A}$ (consumption households), $P3_{S13}$ (consumption government), $P51G$ (gross fixed capital formation), $P5M$ (changes in stocks and inventories), $P6$ (exports) and SD (statistical discrepancy due to chain linking), that is,

$$B1G = -P7 + P3_{S1A} + P3_{S13} + P51G + P5M + P6 + SD. \tag{1}$$

Statistics Netherlands publishes a very detailed tree-structured breakdown into expenditures, as explained in the Appendix (Section 6) on the breakdown in expenditures. The production approach is not considered in this article. Any breakdown of GDP has the same problems with additivity, so for brevity, we only use the above breakdown in this article.

The way we will handle the discrepancies arising from chain linking is by considering them as an extra subseries of the aggregate. It is a series that must be adjusted for seasonal and calendar effects, together with the other subseries of the aggregate. Therefore, the method we developed is also suitable for series where no chain linking takes place, such as current price data and any other set of series where preserving additivity, or at least reducing discrepancies due to adjustments for seasonal and calendar effects, is required.

As we will apply our model to chain volumes of GDP and its expenditures, the total statistical discrepancy after adjustments for seasonal and calendar effects can be divided in two parts, each with its own origin.

2.1. *Discrepancies Arising from Chain Linking*

The statistical discrepancies due to chain linking can be interpreted as the consequence of changes in relative prices of subseries of the aggregate. One can show that the sum of the chain linked expenditures is chain volume with different weights. In each link step, the chain volume of GDP is weighted with the relative values of its expenditures in the previous year, valued at prices of the previous year, whereas the sum of expenditures is weighted with the relative values of the previous year, in reference year prices. So, the statistical discrepancy due to chain linking is the difference between the value of the aggregate valued in previous year prices and the value of the aggregate valued in reference year prices. The discrepancies due to chain linking are therefore zero in the first year after the reference year and tend to be larger the further away they are from the reference year.

The statistical discrepancies due to chain linking typically have a slow moving long-term trendcycle, combined with a strong short-term pattern. The short-term pattern has a seasonal and an irregular component. Figure 1 shows a typical example for total GDP.

The statistical discrepancies due to chain linking are inherent to the way they are defined and should not be corrected, as this would harm the essence of a chain linked volume. The seasonal pattern of the discrepancies can be removed. In theory, the discrepancies due to chain linking could also show calendar effects. However, in the case of the Dutch economic series, they are negligibly small and we choose to ignore them.

The quarter-to-quarter changes of the seasonal adjusted GDP (GDP-SA) are very important results from the economic analysis. Therefore, it makes sense to also calculate the quarter-to-quarter changes of the statistical discrepancy and compare them to GDP-SA as follows:

$$\%SD_t = (SD_t - SD_{t-1})/B1G_{t-1} * 100\%. \tag{2}$$

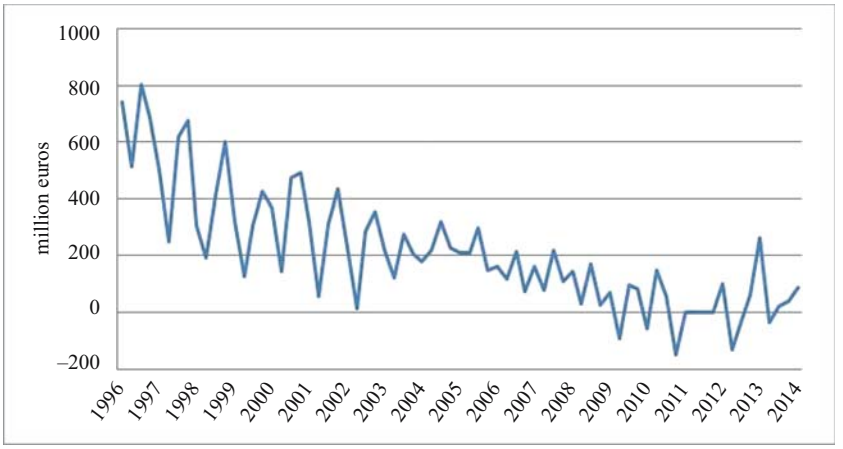

Fig. 1. Statistical discrepancies due to chain linking between Dutch GDP and the sum of the final expenditures (before adjustments for seasonal and calendar effects), reference year = 2010.

When these quarter-to-quarter changes are of similar magnitude or larger than the changes of GDP-SA itself, the analysis of the latter, by breaking it down into components, is severely hampered.

Especially when looking at quarter-to-quarter changes, removing the seasonal pattern can lead to a large reduction in the size of the discrepancies from chain linking. This is shown in Figure 2, where the quarter-to-quarter changes of the discrepancies in percentages of GDP-SA have been adjusted for seasonal effects (i.e., the remaining series represents trend-cycle + irregular). The value of this series is mainly between −0.1% and 0.1%. To put this into perspective, the majority of GDP-SA growth rates are between −0.5% and 0.5%. As can be seen, the seasonal component is by far the largest component of the statistical discrepancy arising from chain linking. Therefore, with ideal seasonal and calendar adjustment, the adjusted statistical discrepancy should have a small influence on the interpretation of the economic growth and its components.

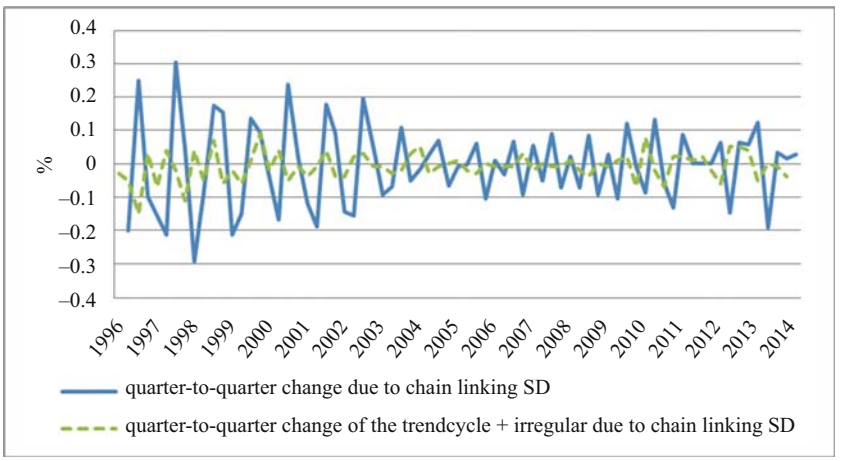

Fig. 2. Univariate seasonal correction of the discrepancies arising from chain linking.

2.2. Discrepancies from Adjustments for Seasonal and Calendar Effects

The second part of the total statistical discrepancies is introduced by the estimation of seasonal and calendar effects. Seasonal and calendar adjustment assumes the following decomposition:

$$y_t = L_t + S_t + \beta x_t + OL_t + SB_t + I_t. \tag{3}$$

Here y_t stands for any of the series appearing in Equation (1), L denotes the trend-cycle, S denotes the seasonal component, βx denotes the regression component with x as an auxiliary variable and β as the regression coefficient, OL denotes additive outliers and level shifts and SB denotes seasonal breaks. Finally, I is an irregular component for the unexplained variation. In this application, the regression component is used to adjust for calendar effects. In general, other regression effects can also be included, but this is not applied here. In a fully consistent adjustment, Equation (1) holds for each of the components in Equation (3). However, when these relations are not explicitly enforced, discrepancies will arise.

The process of seasonal and calendar adjustment consists of a pretreatment phase and the actual seasonal adjustment. In the pretreatment phase, we choose between multiplicative or additive adjustment. In the first case, the original series are logarithmically transformed before decomposition according to Equation (3) is computed. The other parts of the pretreatment phase are adjustments for calendar effects and other regression effects, removal of additive outliers, level shifts and seasonal breaks, and extrapolation of the series in order to apply symmetric filters. The actual seasonal adjustment consists in the application of seasonal and trend filters. In this phase, outlier detection takes place again. After seasonal and calendar adjustment, additive outliers and level shifts are reintroduced into the series. The final adjusted series is therefore equal to:

$$y_t^{SA} = y_t - S_t - \beta x_t - SB_t = L_t + OL_t + I_t. \tag{4}$$

All steps of the process may cause discrepancies:

Logarithmic transformation: Usually, this is done when this yields a better model fit, as, for instance, is current practice in X-13ARIMA-SEATS, see U.S. Census Bureau (2015). For some of our series, this would indeed be the preferred option. However, when multiplicative adjustment is chosen in at least one series, the logarithmic transformation can cause additional discrepancies.

Outlier detection: when each time series is analysed separately for significant outliers, discrepancies may arise when an outlier is significant in one series but not significant or even detectable in another. These may lead to relatively large incidental discrepancies. The outlier detection in both the pretreatment phase and the filtering phase can generate discrepancies. Here, the general term outlier is used for the combination of additive outliers, level shifts and seasonal breaks. A special case is the situation where the seasonal patterns change rapidly. In this case, discrepancies may arise around the period where the rapid change occurs, because these periods are considered to be outliers.

Calendar effects: Estimating the regression coefficients for the calendar effects for each series separately contributes to the discrepancies. The calendar effects in some series are not significantly different from zero (at a 5% significance level). Therefore, it is not

incorporated in the model of these series. This leads to relatively small discrepancies, evenly distributed along the length of the time series.

Extrapolations: The extrapolations are very sensitive to the model choice in X-13ARIMA-SEATS and to outliers at the beginning and end of the time series. This may lead to relatively large discrepancies at the beginning and end of the time series, and is a source of revisions.

Seasonal and trend filters: when different time series are treated with filters of a different length, which is usually the case, some discrepancies will arise in the seasonal components along the full length of the series.

The statistical discrepancy can be calculated by rewriting Equation (1):

$$SD = B1G - (-P7 + P3_{S1A} + P3_{S13} + P51G + P5M + P6). \tag{5}$$

The right-hand side of this equation is called the indirect discrepancy, and the left hand side is called the direct discrepancy. This equation holds not only for the series itself, but in an ideal world, also for each of the components of Equation (3). However, due to the arguments mentioned above, this is not the case for the seasonal component, as is shown in Figure 3 for the period 1996–2014. The solid line is the indirect seasonal component from the chain linked index, calculated as the seasonal component of GDP minus the seasonal components of all other expenditures (right-hand side of Equation 5). The solid line is very different from the seasonal component of the SD (left-hand side of Equation 5). The result is an increase in the quarter-to-quarter change of the discrepancy instead of a significant decrease. Therefore, the analysis of the economic growth and its components is severely hampered.

In a preliminary study, we tried to reduce the inconsistencies by improving the settings in X-13ARIMA-SEATS. We found that a reduction is possible, but reducing them to an

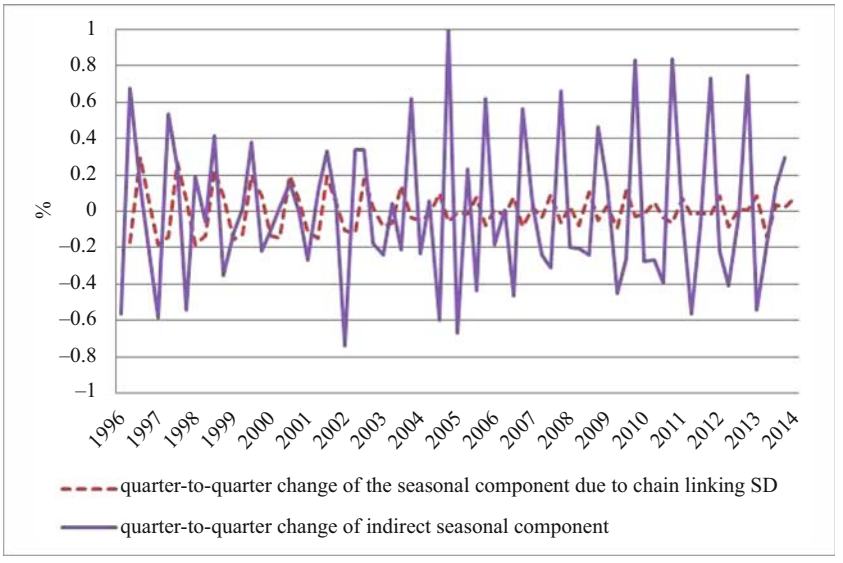

Fig. 3. Seasonal component of the discrepancies arising from chain linking calculated directly and indirectly with conventional univariate seasonal adjustment.

acceptable level or even eliminating them completely and at the same time maintaining sufficient quality of the seasonal and calendar correction does not seem to be possible in this application. The lessons we learned in this study are nevertheless useful for the solution we found, described below. The most important lessons are:

- The largest part of the discrepancies is caused by the pretreatment, especially the treatment of outliers.
- An additive adjustment approach for all series helps to reduce the discrepancies compared to a multiplicative adjustment. Although a multiplicative adjustment is preferred for the series of export, GDP and consumption of households, an additive adjustment is applied. The quality under additive correction is acceptable also for the series where multiplicative correction is preferred. The differences between these approaches are small in practice, at least for the GDP in the Dutch situation.
- The best results could be obtained by an equivalent approach, that is, using the same settings (ARIMA models, filters), modelling outliers in the same periods, additive adjustment for all series, and using the same set of auxiliary variables for all series. This results in a less optimal adjustment according to the Q- and M-diagnostics of X-13ARIMA-SEATS.

The discrepancies are still unacceptably large, even under an optimally chosen equivalent approach and therefore not further implemented in the production of official releases. These findings have led to the conclusion that a multivariate approach is needed in order to reach all three objectives, that is, reducing discrepancies while maintaining univariate quality and avoiding large revisions. This approach is presented in the next section.

3. Multivariate Structural Time Series Model

In this section, a multivariate structural time series modelling approach is developed for the purpose of estimating seasonal effects for an aggregated series and its breakdown in K series in a consistent way. With a structural time series model (STM), a series is decomposed in a trend component, seasonal component, regression components and an irregular component. The model can be extended with other components as cyclic components, or with ARMA components to model autocorrelation beyond these structural components, but this is not applied in the present article. For each component, an appropriate stochastic model is assumed which allows the trend, seasonal, and regression coefficients to be time-dependent. See Harvey (1989) and Durbin and Koopman (2012) for an extensive treatment of structural time series modelling. In multivariate STMs, two or more series are modelled simultaneously, which allows modelling cross-sectional dependency between these series.

3.1. Consistent Seasonal Adjustment with a Multivariate STM

We developed a multivariate STM for quarterly GDP, broken down into a hierarchy according to either the expenditure approach or production approach. Either hierarchy contains multiple levels (see Appendix (Section 6) for the breakdown of expenditures). At every level, there is a statistical discrepancy before seasonal and calendar adjustment (but only if measured in constant prices). The breakdown of GDP into seven subseries

(including the statistical discrepancy) is defined in Equation (1). The time series modelling approach outlined in this section can be applied to each hierarchy of GDP (see Appendix (Section 6) for details). It is important that the consistency between all hierarchical levels of GDP is maintained. This is done repeatedly in a top-down approach. In each hierarchy, restrictions are imposed that ensure that for every subsequent level, all time series components are benchmarked to estimates of the aggregate at the higher level.

Let y_{t+} be the GDP as measured on a quarterly basis. In the first step, the following univariate STM is estimated:

$$y_{t+} = L_t + S_t + \alpha\Delta_t^O + \beta_t x_t + \lambda\Delta_t^L + \gamma_t\Delta_t^S + e_t. \tag{6}$$

The trend-cycle L_t is modelled according to the smooth trend model and the seasonal pattern S_t is modelled using a trigonometric model (Durbin and Koopman, 2012, chap. 3; and supplemental file of this article). Furthermore Δ_t^O is a dummy variable, indicating the period in which an additive outlier occurs, that is,

$$\Delta_t^O = \begin{cases} 1 & \text{for the period } t \text{ where an outlier occurs} \\ 0 & \text{for all other periods} \end{cases} \tag{7}$$

and α denotes the corresponding time-invariant regression coefficient measuring the magnitude of the outlier. In (6), Δ_t^L is a dummy variable indicating the period in which a level shift occurs, that is,

$$\Delta_t^L = \begin{cases} 0 & \text{for all } t \text{ before the period in which a level shift occurs} \\ 1 & \text{for all } t \text{ from (and including) the period in which a level shift occurs} \end{cases} \tag{8}$$

and λ denotes a time-invariant regression coefficient measuring the size of the level shift. A break in the seasonal pattern is modelled with a similar intervention variable:

$$\Delta_t^S = \begin{cases} 0 & \text{for all } t \text{ before the period in which a seasonal break occurs} \\ 1 & \text{for all } t \text{ from (and including) the period of the seasonal break} \end{cases} \tag{9}$$

The magnitude of the seasonal break is measured by γ_t, which is defined as a time-invariant trigonometric seasonal model. This implies that all four quarters have their own break (adding up to zero) which are time invariant. Furthermore x_t denotes the number of working days that is used to model calendar effects in period t, and β_t denotes the corresponding time-dependent regression coefficient modelled as a random walk (Durbin and Koopman 2012, chap. 6). The regression coefficient is allowed to vary over time, since GDP generally increases over time and therefore, also, the size of the working day effect. Finally e_t is a disturbance term for any unexplained variations.

In the general case, multiple additive outliers, level shifts and seasonal breaks are possible, and multiple auxiliary variables may be useful. Then, Equation (6) can be adapted in a straightforward way.

Based on (6), smoothed estimates for total GDP (or the aggregated series of another hierarchy) are obtained. In a second step, the K subseries (without the aggregated series),

represented by a K-dimensional vector $(y_{t1}, \ldots, y_{tk}, \ldots, y_{tK})'$, are modelled by a K-dimensional multivariate STM:

$$y_{tk} = L_{tk} + S_{tk} + \alpha_k \Delta_t^O + \beta_{tk} x_t + \lambda_k \Delta_t^L + \gamma_{tk} \Delta_t^S + e_{tk}, \quad k = 1. \ldots K. \tag{10}$$

The various components in (10) are defined similarly as in Equation (6), but now, for each series $k = 1, \ldots K$, separately. Outliers, level shifts and seasonal breaks may be zero for some series if the analysis shows that they do not occur in a particular series.

To avoid an increase of discrepancies due to seasonal and calendar adjustment, several constraints are imposed on the time series components. These constraints ensure that, for each of these components at each point in time, the value for the aggregate series is exactly equal to the sum of the values of the underlying subseries. Therefore, Equation (10) is applied with the restriction that the sum of the various state variables equals the smoothed values obtained in (6). This is done using the benchmark procedure proposed by Doran (1992). We have constraints for the following components:

- The trend components:

$$L_{t+} = \sum_{k=1}^{K} L_{tk} \tag{11}$$

- The regression coefficients for the working day effects:

$$\beta_{t+} = \sum_{k=1}^{K} \beta_{tk} \tag{12}$$

- The seasonal components:

$$S_{t+} = \sum_{k=1}^{K} S_{tk} \tag{13}$$

- Outliers:

$$\alpha_{+} = \sum_{k=1}^{K} \alpha_{tk} \tag{14}$$

- Level shifts:

$$\lambda_{+} = \sum_{k=1}^{K} \lambda_{tk} \tag{15}$$

- Seasonal breaks:

$$\gamma_{t+} = \sum_{k=1}^{K} \gamma_{tk} \tag{16}$$

With the initially intended seasonal adjustment method, the consistent estimates for the seasonal components S_{tk}, seasonal breaks γ_{tk}, and calendar effects $\beta_{tk} x_t$ are used for

seasonal and calendar adjustment. Since these estimates obey restrictions in Equations (13), (16), and (12), the adjustment procedure does not increase discrepancies.

In order to estimate the multivariate STM described above, it is written in state space form. The state space representation of Equations (6) and (10) is given in the supplement of this article. Next, the Kalman filter is used to obtain optimal estimates for all state variables (see Durbin and Koopman 2012; Harvey 1989). The Kalman filter is a recursive procedure to obtain optimal estimates for the state vector at time t based on the data up to and including time period t. These estimates are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing. Several smoothing algorithms are available in the literature. In this article, the fixed interval smoother is applied, which is a broadly applied smoothing algorithm, and these estimates are referred to as the smoothed estimates. The Kalman filter assumes that the hyperparameters are known, which is generally not the case. Therefore, they are estimated with a maximum likelihood procedure. Finally, we apply diffuse initialization of the Kalman filter for all the state variables.

These models are analyzed with a program that was developed in Oxmetrics (Doornik 2009), using the procedures of Ssfpack 3.0 (Koopman et al. 1999, 2008). Ssfpack is a library of subroutines developed for analyzing (multivariate) STMs. Standard model diagnostics summarized in Durbin and Koopman (2012, chap. 2) are applied to evaluate whether the innovations meet the assumption that they are normally and independently distributed.

Several forms of Model (10) are applied to total GDP and its breakdown in seven series defined by (1). The main differences are in the covariance structures assumed for the disturbance terms of the trend, seasonal component and regression coefficients, varying from full covariance matrices, diagonal matrices and diagonal covariance matrices with equal variances for several series. A general result obtained with these models is that the estimated components, especially the seasonal component, were volatile, and subject to large revisions when data points were added to the time series. From a practical perspective, this is not desirable. In many cases, there are problems with the numerical optimization procedure applied to find a maximum of the likelihood function. To circumvent these issues, an alternative three-step approach is developed to solve problems with large discrepancies. This new approach consists of pretreatment based on a multivariate STM that ensures that the components that contributed most to the discrepancies are consistent. The seasonal adjustment itself is carried out with X-13ARIMA-SEATS. A seasonal adjustment approach based on a multivariate STM is left as a topic for further research.

3.2. Multivariate STM for Pretreatment

The following procedure, which combines the advantages of a multivariate approach and the robustness of conventional univariate seasonal adjustment with X-13ARIMA-SEATS, is proposed as a practical solution for handling discrepancies. In a first step, the univariate STM (6) is applied to the aggregated series. Then, the multivariate STM (10) is applied to the K subseries with restrictions (12), (14), (15), (16), and diagonal covariance matrices for the disturbances of the trends and the seasonal components.

These results are used to remove additive outliers, level shifts, seasonal breaks, and calendar effects from the series. In a second step, X-13ARIMA-SEATS via JDemetra+ (Grudkowska 2015) is used for the extrapolation of the series and to obtain seasonally adjusted series by applying trend and seasonal filters. After that, the additive outliers, level shifts, and calendar effects removed in the first step are reintroduced in the series. In a third step, multivariate benchmarking is applied to eliminate any remaining discrepancies.

The final seasonally adjusted series are the seasonally adjusted series plus the level shifts and additive outliers removed in the pretreatment. Seasonal breaks are not added back to the series, because the purpose of seasonal adjustment is to remove seasonal patterns.

Effectively, this means that the multivariate model is only used for pretreatment, comparable to this step in X-13ARIMA-SEATS (except for the extrapolation, which is done by X-13ARIMA-SEATS in our approach). The advantage of carrying out this pretreatment with a multivariate STM is that calendar effects, additive outliers, level shifts and seasonal breaks are fully consistent between the subseries and the aggregate. To this end, we used the same regressors for all series, and outliers were modelled consistently, as described in the next subsection. Since these components obey restrictions (12), (14), (15) and (16), we obtained a close-to-optimal result, which is more stable than the approach based on estimating consistent seasonal effects with the multivariate STM. An empirical result of this approach is that if pretreatment yields small discrepancies, then most likely the final seasonal adjustment will not increase these discrepancies by much, see Subsection 2.2.

The increase in discrepancies due to univariate seasonal adjustment with X-13ARIMA-SEATS is minimized if the same filter length is used for all series. We chose a short seasonal filter, since the seasonal pattern changes quite rapidly. This had only a slight effect on the quality of univariate seasonal adjustment. The ARIMA model used for extrapolation was determined for the aggregate and applied for each series in the breakdown. This procedure results in seasonally adjusted series that have only very small discrepancies. In order to remove these, a multivariate benchmarking procedure was applied.

3.3. Consistent Outlier Detection

Three types of outliers are distinguished: additive outliers, level shifts and seasonal breaks. As outliers can be much larger than the seasonal and calendar effects, their influence can be very large. Detecting and modelling them is crucial for achieving good seasonal adjustment. In some cases, more than one outlier is needed to model the economic events in a short period of time. On the other hand, it is difficult to find the optimal combination of outliers and avoid overfitting of the series. An observation is that not using the right combination of outliers results in a serious deterioration of quality diagnostics of X-13ARIMA-SEATS. Another consequence is that the estimates of all state variables can become unstable.

For every significant outlier in one of the series, there must be a counterpart in one or more of the other series to achieve consistency. These counterparts are not necessarily

significant, and therefore difficult to detect. Sometimes there are substantive economic explanations for the occurrence of outliers that are helpful in choosing the type of outlier, timing, the counterparts and in some cases, even the size of the outliers. When the size of an outlier can be determined from statistical/economic analysis, the outlier can be manually removed from the series and does not need to be modelled in the multivariate STM. This results in a more parsimonious model, which is therefore preferred. When the size of the outliers cannot be determined, we model the outliers in the STM.

For around 50% of the outliers, this additional information is not available and outliers are detected in a model selection process. Outliers are detected using the automatic detection in the pretreatment phase of X-13ARIMA-SEATS. This method is based on a RegARIMA model. These outliers are modelled in the multivariate STM. Additional outliers are detected by considering the residuals of the STM using a disturbance smoother (Harvey and Koopman 1992). For all residuals with a t-value larger than 2.5, an outlier and its counterparts are added to the set.

The set of outliers that is modelled explicitly in the pretreatment phase is removed from the series. The pretreated series become the input for the seasonal adjustment phase. Nevertheless, the seasonal adjustment phase in X-13ARIMA-SEATS can detect additional outliers. As these decisions are made for each series separately, they will again lead to statistical discrepancies. Therefore, we must model these events with additional outliers in the pretreatment phase. In order to reduce the number of outliers detected in the seasonal adjustment phase, we increased the critical value that controls whether an observation is classified as an outlier. All outliers that were detected above this level were added to the set of outliers modelled in the pretreatment phase. This process is iterated until no new outliers are detected.

4. Results

In this section, we apply the seasonal and calendar adjustment approach obtained with standard X-13ARIMA-SEATS (old method) and the improved method proposed in Section 3 to the cycle of releases of the quarterly GDP and its breakdown in components according to (1) in an annual estimation cycle and compare the results obtained under both approaches. The old method refers to the application of X-13ARIMA-SEATS with settings that conform to Eurostat guidelines (Eurostat 2015), as applied in the production process. This means that for part of the series the logarithmic transformation is applied. The set of outliers under the old method is different from the set under the improved methods.

The quarterly GDP figures are produced twice: 45 days after the end of a quarter, a flash estimate is published. Then, 85 days after the end of the quarter a new regular estimate is published, based on more complete data sources. When the regular estimate of the fourth quarter is published in March, the figures for the first three quarters are revised again.

The quarterly figures are revised three more times after that: for each new annual estimate, the quarterly figures are adapted such that the four quarters add up to the new annual figure. This happens for the first time in June and for the second time one year later, when the final annual figures are published. Finally, one year after this, the quarterly figures are revised one more time without changing the annual results. Furthermore, every

time a new quarter is added, the seasonal adjustment procedure is applied to the entire time series, potentially affecting all quarters. However, normally, revisions of seasonal adjustments to earlier figures are small. Since 2016, the figures are revised twice because the process has been accelerated. This implies that the second revision is the final one, where only quarterly figures are adjusted.

This means that once a year, at the time that the regular estimate for the first quarter (1r) is made, large changes are made to the (unadjusted) time series. Therefore, it is necessary to derive new settings for the seasonal and calendar adjustment at this point in time every year. The annual estimation cycle starts with the regular estimate for the first quarter, and the derived settings are then used for all subsequent estimates of the annual estimation cycle. The 1r estimate is followed by the first (flash) estimate of the second quarter (2f) and the second estimate of the second quarter (2r). This scheme is continued until the second estimate of the fourth quarter (4r). The first estimate of the first quarter of the current year comes before the large updates of 1r and therefore, is also part of the same cycle. It is called 5f, to emphasize that the settings of the previous year are applied.

Below, the old method and the improved method (without the final benchmarking step) are compared according to the three quality criteria described in Section 1:

- The statistical discrepancy between the seasonally adjusted GDP and its components.
- The standard quality diagnostics of X-13ARIMA-SEATS: M1 to M11 and Q.
- The revisions of the published results between the subsequent estimations.

We use data from the time period 1996–2014 for the computations in this section.

4.1. *The Statistical Discrepancy due to Seasonal and Calendar Adjustment*

In this subsection, we discuss the discrepancies added by the seasonal and calendar adjustment process. This process estimates the seasonal components of each of the series. In both the old approach and the improved approach, the estimated seasonal components of all subseries do not add up to the seasonal component of GDP, and result in a residual:

$$\Delta = S_{B1G} + S_{P7} - S_{P6} - S_{P3_{S1A}} - S_{P3_{S13}} - S_{P51G} - S_{P5M} - S_{SD}$$

In Table 1, we compare the added statistical discrepancy of the old method and the improved method, by taking the relative added statistical discrepancy computed as a

Table 1. Average and maximum absolute discrepancies due to seasonal adjustment for the year 2014.

	Old method		Improved method	
	Avg %	Max %	Avg %	Max %
1r	0.331	1.01	0.001	0.01
2f	0.329	1.00	0.001	0.01
2r	0.321	1.00	0.001	0.01
3f	0.321	0.99	0.001	0.01
3r	0.320	0.99	0.001	0.01
4f	0.322	0.98	0.001	0.01
4r	0.330	0.99	0.001	0.01
5f	0.313	0.99	0.000	0.01

percentage change from seasonally adjusted GDP:

$$\left|\frac{\Delta}{B1G^{SA}}\right| * 100\%$$

Table 1 presents the average and maximum of this difference over the entire time series.

Table 1 shows that a significant reduction in statistical discrepancy due to seasonal and calendar adjustment can be achieved by using the improved method. With the old method, interpretation of GDP growth was, on average, hampered by the discrepancy by 0.3%, with a maximum of 1%, while, with the new method, the disturbance is negligible.

4.2. The Standard Diagnostics of X-13ARIMA-SEATS

Software of the X-11-family summarizes the quality of the seasonal and calendar adjustment with M1 to M11 and a Q-diagnostics. These diagnostics value different aspects of the seasonally adjusted series. For the meaning of the values, see the supplemental file or Ladiray and Quennville (2001). These statistics vary between 0 and 3. Values smaller than 1 are to be preferred, however are not always achievable due to characteristics of the series. The lower the value, the better. Tables 2 and 3 present the diagnostics of the seasonal adjustment with the old method and the improved method of estimate 1r.

In Table 2, eight diagnostics are between 1 and 2 and two of them are above 2. This shows that the quality of the seasonal and calendar adjustment is not always satisfactory, but further substantial improvements are not possible using traditional methods. With the new method, the results have improved to six diagnostics above 1 and none above 2. On the other hand, the new method has fewer quality diagnostics with very small values. On average, the quality improves slightly. This is due to the improved analysis of the outliers. The multivariate pretreatment of the new method results in fewer quality diagnostics with very high and very low values.

Table 2. Quality of seasonally adjusted estimate 1r (for the year 2014) with old method (values > 1 are bold).

	Import ($P7$)	Consumption HH ($P3_{S1A}$)	Consumption govern ($P3_{S13}$)	Cap. form. ($P51G$)	Stocks ($P5M$)	Export ($P6$)	GDP ($B1G$)
M1	0.49	0.42	0.00	0.09	0.33	0.10	0.05
M2	0.64	0.00	0.00	0.06	0.19	0.05	0.02
M3	0.13	0.00	0.00	0.31	0.39	0.00	0.00
M4	0.18	**1.16**	0.95	**1.05**	0.84	0.84	0.84
M5	0.24	0.20	0.20	0.20	0.41	0.20	0.20
M6	0.12	0.44	**1.00**	0.16	0.64	0.52	0.22
M7	0.38	0.39	0.16	0.13	0.32	0.24	0.06
M8	0.91	0.71	0.51	0.53	**1.12**	0.60	0.17
M9	0.54	0.64	0.26	0.14	0.67	0.34	0.06
M10	**1.33**	**1.04**	0.38	0.54	**2.30**	0.98	0.17
M11	**1.30**	**1.02**	0.22	0.44	**2.30**	0.97	0.17
Q	0.46	0.45	0.22	0.29	0.64	0.34	0.16

Table 3. Quality of seasonally adjusted estimate 1r (for the year 2014) with improved method.

	Import ($P7$)	Consumption HH ($P3_{S1A}$)	Consumption govern ($P3_{S13}$)	Cap. form. ($P51G$)	Stocks ($P5M$)	Export ($P6$)	GDP ($B1G$)
M1	0.33	0.61	0.02	0.27	0.73	0.18	0.04
M2	0.17	0.02	0.03	0.18	0.55	0.08	0.02
M3	0.30	0.17	0.27	0.86	**1.10**	0.03	0.00
M4	0.95	0.51	0.51	0.40	0.18	**1.05**	0.40
M5	0.20	0.20	0.20	0.60	0.95	0.20	0.20
M6	0.08	**1.32**	0.74	**1.41**	0.81	0.31	0.16
M7	0.30	0.19	0.24	0.14	0.19	0.18	0.10
M8	0.82	0.72	0.51	0.68	**1.02**	0.60	0.33
M9	0.45	0.13	0.26	0.19	0.21	0.25	0.22
M10	0.64	0.52	0.36	0.78	**1.23**	0.38	0.22
M11	0.32	0.40	0.24	0.17	0.48	0.16	0.12
Q	0.40	0.30	0.24	0.39	0.61	0.28	0.14

The quality of GDP is almost the same as before, despite the fact that in the multivariate approach the excellent univariate seasonal and calendar adjustment is slightly disturbed by the other series. The quality of the gross fixed capital formation ($P51G$) deteriorates because M5 worsens due to the larger level shift in 2009-Q1 (resulting in less trend) in the multivariate case compared to the univariate case. M1 and M2 deteriorate under the new method because this series contains a larger irregular component. This is the result of using larger critical values for detecting outliers in X-13ARIMA-SEATS in the seasonal adjustment phase. This also affects M6 and M8, resulting in larger arbitrary changes of the seasonal component. The results for M9 to M11 are greatly improved. This is caused by the modelling of the seasonal outliers. In both methods, the sum of the four quarters of a seasonal outlier adds up to zero. However, with the improved method, a seasonal outlier has a different magnitude in every quarter, while with the old method, the outlier is determined in one quarter and the three other quarters have a third of its opposite magnitude. A disadvantage of the improved method is that it uses three quarters in the time series to determine the outlier, while the old method only uses one degree of freedom.

Remarkably large differences in diagnostics are found for M4 for the import ($P7$) and the export ($P6$): both deteriorate. Further analysis showed that M4 could be improved by adding an extra seasonal break in 2003, which becomes more pronounced due to the seasonal break of 2008. However, this was unknown during the implementation of the new method for the seasonal adjustment of the Dutch quarterly national accounts.

Table 4 presents the difference in overall quality (as measured by the Q-diagnostic) between the two methods for all eight estimates. Negative values (in bold) relate to an improvement by using the new method, positive values relate to a deterioration. Both methods display an almost constant difference in quality during the annual cycle.

Table 4. Difference in Q-diagnostic between old and new method for seasonal adjustment, for the year 2014.

	Import ($P7$)	Consumption HH ($P3_{S1A}$)	Consumption govern ($P3_{S13}$)	Stocks ($P5M$)	Cap. form. ($P51G$)	Export ($P6$)	GDP ($B1G$)
1r	**−0.06**	**−0.15**	0.02	0.10	**−0.02**	**−0.07**	**−0.01**
2f	**−0.08**	**−0.10**	0.03	0.10	**−0.07**	**−0.07**	**−0.02**
2r	**−0.07**	**−0.10**	0.03	0.10	**−0.07**	**−0.06**	**−0.01**
3f	**−0.08**	**−0.11**	0.02	0.13	**−0.06**	**−0.07**	**−0.01**
3r	**−0.08**	**−0.11**	0.02	0.13	**−0.06**	**−0.07**	**−0.01**
4f	**−0.07**	**−0.12**	0.00	0.16	**−0.04**	**−0.08**	**−0.01**
4r	**−0.06**	**−0.16**	0.00	0.09	**−0.03**	**−0.08**	**−0.02**
5f	**−0.05**	**−0.14**	0.01	0.12	**−0.03**	**−0.06**	**−0.02**

4.3. Revisions

In this section, revisions of the quarter-to-quarter growth in %-point are investigated under the old method and the improved method.

The quarter-to-quarter growth is defined as

$$\hat{\theta}_t = \frac{y_t^{SA} - y_{t-1}^{SA}}{y_{t-1}^{SA}} \cdot 100\%, \tag{17}$$

where y_t^{SA} denotes the seasonally adjusted figures. This is computed for the GDP and the variables of the breakdown.

The revisions are split into two types; the first are due to the updates from flash to regular estimate:

$$R_1 = \left(\sum_{t=2}^{4} \left| \hat{\theta}_{t|t}^{f} - \hat{\theta}_{t|t}^{r} \right| \right). \tag{18}$$

with $\hat{\theta}_{t|T}^{f}$ and $\hat{\theta}_{t|T}^{r}$ the flash and regular estimates of the growth rates, see formula (17), for quarter t based on the time series up to and including quarter T.

Note that the first quarter ($t = 1$) is excluded in R_1 since in the regular estimate of this quarter, the information on an annual basis is added, which causes large revisions.

The second type of revisions is due to adding the flash estimate of a new quarter:

$$R_2 = \left(\sum_{t=1}^{4} \left| \hat{\theta}_{t|t}^{r} - \hat{\theta}_{t|t+1}^{f} \right| \right).$$

The average revision of the last quarter is therefore:

$$\frac{1}{7}(R_1 + R_2) \tag{19}$$

which is presented in Figure 4.

Similarly, Figure 5 presents the average absolute revisions of the quarter-to-quarter growth in %-point over eight estimates (i.e., seven differences) over the last year

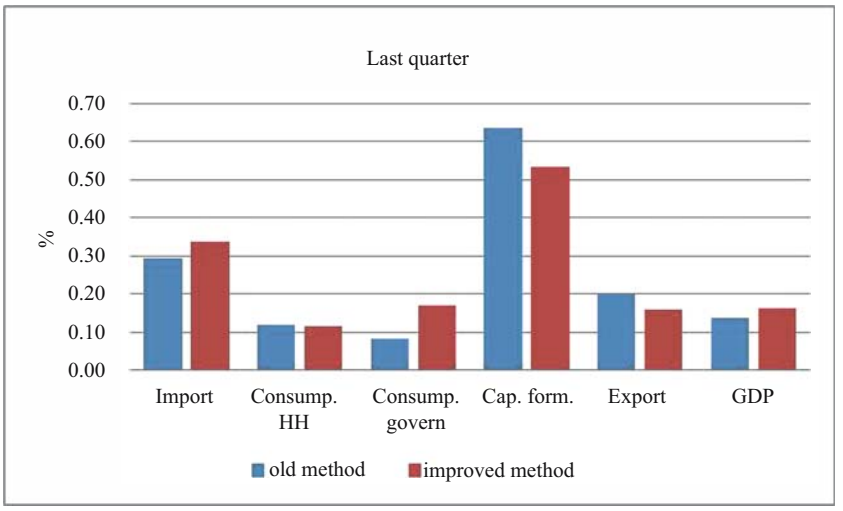

Fig. 4. Average revision of quarter-to-quarter growth in 2014 defined by equation (19).

averaged per quarter:

$$\frac{1}{28}\sum_{j=0}^{3}(R_{1j}+R_{2j}) \tag{20}$$

$$\text{with} \quad R_{1j}=\left(\sum_{t=2}^{4}\left|\hat{\theta}^{f}_{t-j|t}-\hat{\theta}^{r}_{t-j|t}\right|\right), \text{j}=0,1,2,3$$

$$\text{and} \quad R_{2j}=\left(\sum_{t=1}^{4}\left|\hat{\theta}^{r}_{t-j|t}-\hat{\theta}^{f}_{t-j|t+1}\right|\right), \text{j}=0,1,2,3$$

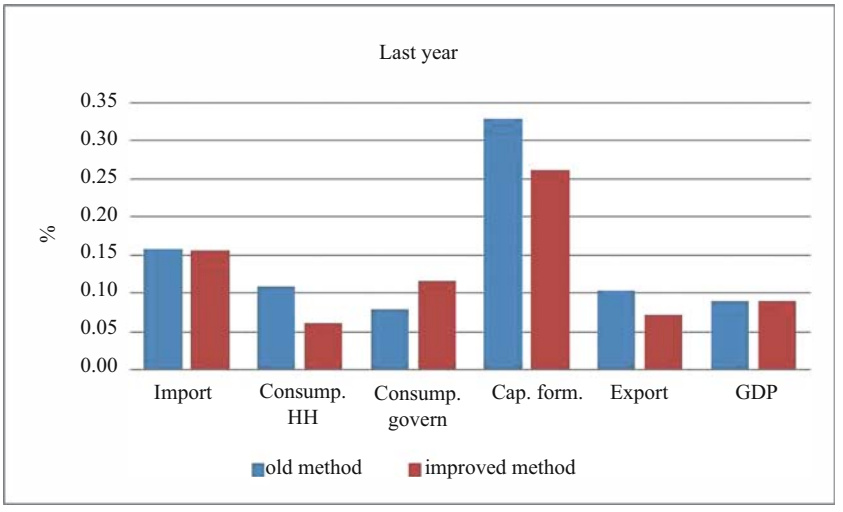

Fig. 5. Average revision of annual growth in 2014 defined by equation (20).

Series SD (statistical discrepancy due to chain linking) and P5M (changes in stocks) are both fluctuating around zero. Therefore, both can have huge growths in %-points because of small absolute values, resulting in huge revisions of the growth. As a consequence, they are left out of the analysis. The figures show that the size of the revisions of the old method and the improved method are almost equal. For consumption government a small deterioration is observed, but the revisions are still very small. The deterioration is caused by the time-varying seasonal pattern, which is picked up faster under the improved method due to the use of shorter seasonal filters. A reduction of the revisions was not expected in advance, as adding or changing observations at the end of the series gives new information about trend-cycle and seasonal component. Revisions are therefore inherent to seasonal and calendar adjustment.

5. Conclusion

Quarterly figures about GDP with a breakdown in K underlying subseries for, for example, expenditures or industries, are produced by national statistical institutes to measure and analyze economic growth. Two factors are responsible for discrepancies between the sum of the underlying K subseries and the total GDP. The first factor arises due to the process of chain linking, which means that series of volume growth rates are expressed in constant price levels. Since the annual changes of these price levels differ between the series, statistical discrepancies between the sum of the underlying series and total GDP arise. The first factor does not arise if the estimate is in current values. The second factor arises after adjusting for seasonal and calendar effects using the standard approach based on X-13ARIMA-SEATS. In the Netherlands, since 2009, the size of these discrepancies has often been larger than the growth rates of GDP itself and hampers the interpretation of these figures.

Several intuitive approaches to avoid discrepancies are available in the literature, such as the indirect approach and multivariate benchmarking. A major drawback of the first approach is that official figures about GDP are derived from the most detailed breakdown, which contains the largest fluctuations, while the most reliable estimates at the aggregated level are not used. Benchmarking is appropriate if the discrepancies are modest. In the Dutch application, the discrepancies are large, and benchmarking introduces a residual seasonal effect in the adjusted series.

In this article, an alternative approach based on a multivariate structural time series model is considered. The most intuitive approach is to construct a $K + 1$ dimensional structural time series model for GDP and its breakdown in K subseries. The model contains explicit constraints on the state variables to ensure that trend, seasonal effects, calendar effects and outliers in GDP are equal to the sum of the K subseries of these components. In this way, available series are consistently modelled and a two-stage approach is avoided. Nevertheless, the results obtained with this approach are not satisfactory since the estimated seasonal effects are too volatile. Furthermore, we observed numerical problems with the maximum likelihood procedure for the hyperparameters and the revisions were too large. Solving these problems is left as a topic for further research.

As an alternative, a multivariate structural time series model with consistency restrictions on the additive outliers, level breaks, seasonal breaks and calendar effects (derived from a univariate model applied to the aggregated series) is only used to eliminate these effects from the observed series. Subsequently, X-13ARIMA-SEATS is used for seasonal adjustment of all series. This reduces the inconsistencies remarkably. Finally, a multivariate benchmarking is applied to restore consistency in the adjusted series, and the additive outliers, level breaks, and calendar effects are added to the adjusted series. With this pretreatment approach, a significant reduction of the statistical discrepancies is achieved, whereas the quality of the adjustment in terms of the standard X-13ARIMA-SEATS quality measures is maintained or even improved for some series. In June 2015, this approach was implemented in the production of Dutch official statistics on economic growth.

When the numerical problems with the complete multivariate approach can be solved, comparing results of both approaches can give some insights into the influence of the approaches on the estimates.

The approach considered in this article is generic and applies to many other applications at national statistical institutes. Therefore, it is worthwhile to further improve the $K + 1$ dimensional structural time series model, where consistent seasonal effects are directly estimated with the structural time series model.

6. Appendix: Detailed breakdown of GDP

Figure 6 summarizes the breakdown of GDP according to the expenditure approach. The official tables published by Statistics Netherlands are actually more detailed. Not presented here is the further breakdown of gross fixed capital formation. Each hierarchy of this breakdown is consistently corrected for seasonal and calendar effect using the top-down approach described in Subsection 3.2.

In each branch, there is a time series marked *SDx*. These are the discrepancies arising from chain linking. They only occur in constant price data. The figure illustrates that the breakdown comprises GDP itself and 20 subseries, complemented by nine different series for the discrepancies arising from chain linking (one for each branch of the tree). A similar breakdown tree of GDP is used for value added by industry (the production approach). This tree consists of six branches, and comprises 20 subseries for branches of industry and, of course, six series for discrepancies arising from chain linking.

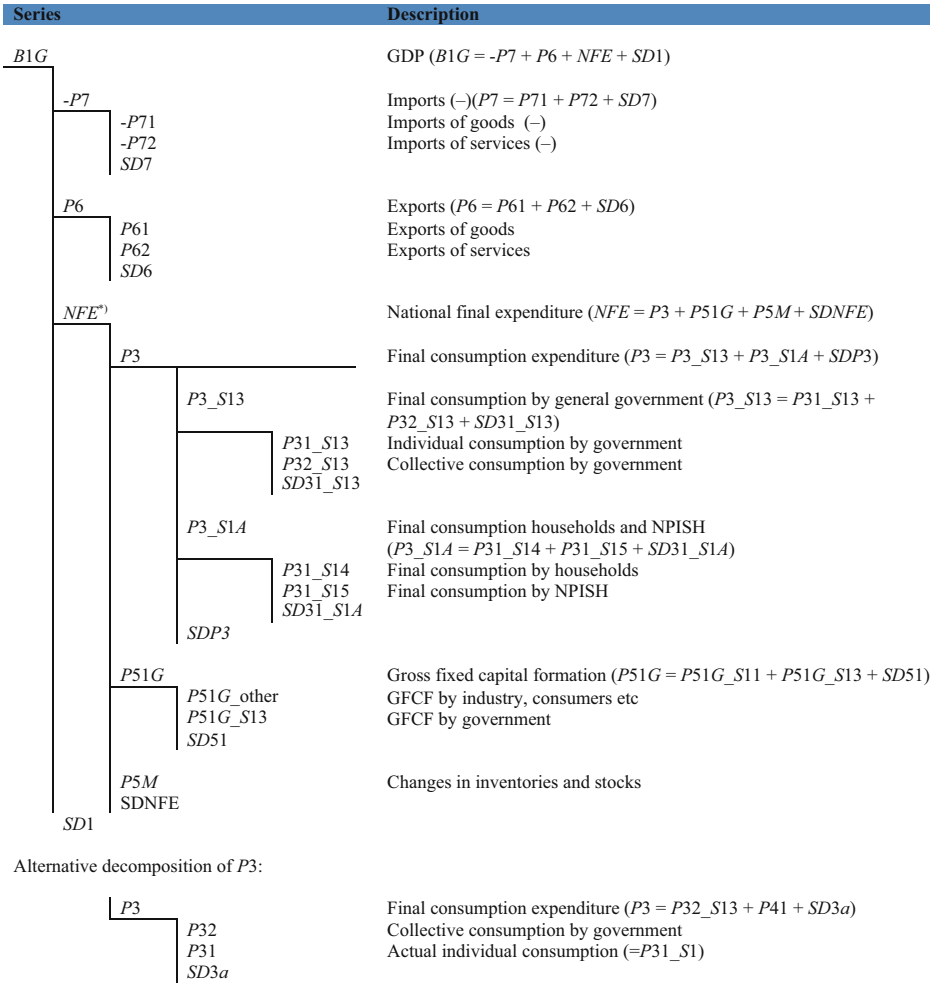

Fig. 6. Breakdown of GDP in components.

7. References

Bloem, A., R. Dippelsman, and N. Maehle. 2001. *Quarterly National Accounts Manual–Concepts, Data sources and Compilation*. Washington D.C. IMF. Available at: https://www.imf.org/external/pubs/ft/qna/2000/Textbook/ (accessed January 2019).

Di Fonzo, T. and M. Marini. 2011. "Simultaneous and Two-step Reconciliation of Systems of Time Series: Methodological and Practical Issues." *Journal of the Royal Statistical Society C (Applied Statistics)* 60. Part 2, 143–164. Doi: https://doi.org/10.1111/j.1467-9876.2010.00733.x.

Doornik, J.A. 2009. *An Object-oriented Matrix Programming Language Ox 6*. London: Timberlake Consultants Press.

Doran, H.E. 1992. "Constraining Kalman Filter and Smoothing Estimates to Satisfy Time Varying Restrictions." *Review in Economics and Statistics* 74: 568–572. Doi: https://doi.org/10.2307/2109505.

Durbin, J. and S.J. Koopman. 2012. *Time Series Analysis By State Space Methods* (Second Edition). Oxford: Oxford University Press.

Eurostat. 2015. *ESS Guidelines on Seasonal Adjustment* (2015 Edition). Luxembourg: European Union. Available at: https://ec.europa.eu/eurostat/cros/content/ess-guidelines-seasonal-adjustment-2015-edition_en (accessed January 2019).

Grudkowska, S. 2015. *JDemetra and User Guide*. Warsaw: National Bank of Poland, Department of Statistics. Available at: https://ec.europa.eu/eurostat/cros/system/files/jdemetra_user_guide.pdf (accessed January 2019).

Harvey, A.C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, J.A. and S.J. Koopman. 1992. "Diagnostic Checking of Unobserved Components Time Series Models." *Journal of Business and Economic Statistics* 10: 377–389. Doi: https://doi.org/10.1080/07350015.1992.10509913.

Koopman, S.J., N. Shephard, and J.A. Doornik. 1999. "Statistical Algorithms for Models in State Space Form Using Ssfpack 2.2." *Econometrics Journal* 2: 107–160. Doi: https://doi.org/10.1111/1368-423X.00023.

Koopman, S.J., N. Shephard, and J.A. Doornik. 2008. *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*. London: Timberlake Consultants Press.

Ladiray, D. and B. Quenneville. 2001. *Seasonal Adjustment with the X-11 Method*. New York: Springer-Verlag.

Quenneville, B. and S. Fortier. 2006. "Balancing Seasonally Adjusted Series as a Complement to the Direct and Indirect Approaches to Seasonal Adjustment." *Proc. Bus. Econ. Statist. Sect. Am. Statist. Ass.*: 1118–1125.

U.S. Census Bureau. 2015. *X-13ARIMA-SEATS Reference Manual, Version 1.1*. Washington, D.C: U.S. Census Bureau. Available at: https://www.census.gov/ts/x13as/docX13AS.pdf (accessed January 2019).

Received July 2017
Revised March 2018
Accepted August 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 31–65, http://dx.doi.org/10.2478/JOS-2019-0003

Is the Top Tail of the Wealth Distribution the Missing Link between the Household Finance and Consumption Survey and National Accounts?

Robin Chakraborty[1], *Ilja Kristian Kavonius*[2], *Sébastien Pérez-Duarte*[3], *and Philip Vermeulen*[3]

The financial accounts of the household sector within the system of national accounts report the aggregate asset holdings and liabilities of all households within a country. In principle, when household wealth surveys are explicitly designed to be representative of all households, aggregating these microdata should correspond to the macro-aggregates. In practice, however, differences are large. We first discuss conceptual and generic differences between those two sources of data. Thereafter, we investigate missing top tail observation from wealth surveys as a source of discrepancy. By fitting a Pareto distribution to the upper tail, we provide an estimate of how much of the gap between the micro- and macrodata is caused by the underestimation of the top tail of the wealth distribution. Conceptual and generic differences, as well as missing top tail observations, explain part of the gap between financial accounts and survey aggregates.

Key words: Financial accounts; HFCS; wealth inequality; Pareto distribution; households.

1. Introduction

Household wealth surveys provide detailed information on the value of assets and liabilities held by individual households within a country. The financial accounts (FA) of the household sector within the system of national accounts (SNA) report the value of aggregate asset holdings and liabilities of all the resident households. In principle, when household wealth surveys are explicitly designed to be representative of all resident households in the country, aggregating these microdata should correspond to the macro-aggregates. In practice, however, differences are large, where usually the value of the aggregated microdata is below the macro-aggregates. This fact has given birth to a new literature (Antoniewicz 2000; Kavonius and Törmälehto 2010; Henriques and Hsu 2014;

[1] Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, DE-60431 Frankfurt am Main, Germany. Email: Robin.chakraborty@bundesbank.de
[2] University of Helsinki, P.O. Box 3, FI-00014, Finland. Email: ilja.kavonius@helsinki.fi
[3] European Central Bank, DE-60640 Frankfurt am Main, Germany. Emails: sebastien.Perez_Duarte@ecb.int and philip.vermeulen@ecb.europa.eu

Acknowledgments: This article uses data from the Eurosystem Household Finance and Consumption Survey. The article has benefited from the discussions of the Expert Group on Linking Macro and Micro Statistics for the Household Sector (EG-LMM). We would like to thank Peter van de Ven (OECD) for discussing the paper at the IARIW 34th General Conference and for his valuable comments. Additionally, we would like to thank Caroline Willeke, Prasada Rao, Henning Ahnert, two anonymous referees and the Associate Editor for their valuable comments. This article should not be reported as representing the views of the Deutsche Bundesbank or the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the Deutsche Bundesbank or the ECB. Any remaining errors are solely ours.

Kavonius and Honkkila 2013; Andreasch et al. 2013; and Dettling et al. 2015), which attempts to understand the striking differences observed between aggregates produced by household wealth surveys and those reported in the financial accounts. This article contributes to this emerging literature.

The reconciliation of household wealth surveys with FA data is an important issue for a number of reasons. First, household wealth surveys have been combined with FA data (and with other administrative data) to analyse the evolution of wealth inequality. In a recent paper, Bricker et al. (2016a) show that calibrating the US Survey of Consumer Finances data to the FA substantially affects the top shares of the wealth distribution. This helps to explain why Saez and Zucman (2016), who also use the FA and combine it with tax records, obtain higher and faster rising shares.

Inequality is high on the political and economic research agenda. Stiglitz et al. (2009) and Piketty (2014) illustrate the importance of distributional information of wealth in analysing economic progress. Central banks are also increasingly interested in the distributional issues, as these have been recognised to interact with monetary policy. For instance, the IMF/FSB report to the G20 Finance Ministers and Central Bank Governors' data gap initiative emphasised, in particular, a need for including distributional information in macrodata. Tissot (2015) discusses the G20 Data Gap Initiative, the benefits of collecting microdata and its interest for macroprudential and monetary policies. When aggregate wealth from wealth surveys differs substantially from macro-aggregates, the inequality measured using such surveys can become questionable.

Second, several European and international groups have been established with the underlying motivation to include distributional measures in the SNA, as well as having timely distributional data. Survey information is likely to be used as one input. However, before such survey information can be used satisfactorily, the observed differences with the FA have to be understood. Our work can be seen in light of the following initiatives. In the beginning of 2016, the European Central Bank (ECB) established an Expert Group on Linking Micro and Macro Household Data (EG-LMM). The focus is on linking FA balance sheet data with the HFCS. The results in Section 2 of this article benefit from the discussions of that group. Similar kind of work has been done in the United States (see e.g., Dettling et al. 2015; Henriques and Hsu 2014 and Antoniewich 2000 for comparisons between the Flows of Funds and the Survey of Consumer Finances). While the scientific discussion about Distributional National Accounts in a sense of *national income* (see for example, the work by Piketty et al. 2018), is more advanced, work about Distributional National Accounts in the sense of wealth is very limited so far (see e.g., Alvaredo et al. 2016 and Alvaredo et al. 2017).

While wealth surveys are one distinct source for analysing wealth inequality, research has shown that the upper parts of the wealth distribution are often missing in household wealth surveys (see, e.g., Bach et al. 2015; Eckerstorfer et al. 2016, and Vermeulen 2016, 2018). As the wealth distribution is highly skewed and these upper parts own significant shares of total wealth, this leads to an underestimation of aggregate wealth compared to the FA. The main contribution of this article is to provide estimates on how much of the gap between household wealth surveys and the FA is caused by the underrepresentation of the top tail of the wealth distribution in surveys.

We use the first wave of the Household Finance and Consumption Survey (HFCS) and FA data from Austria, Germany, France, Spain and Finland. This choice of countries is

determined by the need to combine three sources of data: the HFCS data, the FA data and extraneous data that allow us to estimate the top. We use the Forbes billionaires list as such extraneous data.

We first discuss the conceptual linkages and generic statistical differences between the HFCS and the FA. Although both are designed to capture the components of wealth of households, conceptual and statistical differences imply that any comparison has its limitations. We focus on how financial assets are captured in both sources (and leave real assets for future study). First, we do a naïve comparison, where we ignore these conceptual and statistical differences, of total financial assets in the HFCS and the FA. Such a naïve comparison indicates serious differences in the magnitudes between the micro- and the macro-aggregates. Second, we attempt to reconcile the data from HFCS and FA by developing what we call "adjusted concepts of financial assets", which have more comparability between the two data sources. Here we follow the line of work by Kavonius and Törmälehto (2010). We find that gaps become smaller, but are still substantial using these adjusted concepts. Finally, we focus on the wealthiest households and add a Pareto tail to the household wealth surveys to allow for the missing wealthy. We estimate how much of the gap can be attributed to this group. We find that, especially for countries doing no oversampling or having a less effective oversampling strategy, adding a Pareto tail can explain a significant part of the micro-macro gap, while for countries having a more effective oversampling strategy, for example based on taxable wealth, adding a Pareto tail explains less of the gap.

To estimate a Pareto tail, we follow the procedures in Vermeulen (2018) and use three different methods. The estimation method of the Pareto tail is of importance. Using the regression method, including the Forbes data, yields the highest estimates for the tail and can explain more of the micro-macro gap, while using other methods (pseudo maximum likelihood method and the regression method without the Forbes) explains much less. Although including the Forbes data increases the tail significantly, in the cases where countries used an effective oversampling strategy (Spain and France) the micro-macro gap is affected much less. This crucially depends on the weight allocated to the tail in the survey, which is much less in these countries.

The rest of the article is organised as follows: Section 2 analyses the generic statistical and conceptual differences between the two sources. Based on this analysis, we develop two different adjusted concepts for financial assets with the intention on basing the comparison only on those financial instruments that are included in both sources and that are conceptually comparable. For the two adjusted concepts, we indicate the differences between the HFCS and the financial accounts. The third section focuses on the methodology used to estimate the tail of the wealth distribution based on Vermeulen (2018). Finally, in Section 4 we analyse how the estimated tail based on the Pareto distribution changes the remaining gaps for one of the adjusted concepts developed in Section 2. The final section concludes.

2. The Household Finance and Consumption Survey and the Financial Accounts: How Are They Related?

We use survey data from the first wave of the HFCS for Austria, Germany, France, Spain and Finland. The HFCS is a triennial survey that provides individual household data on the

components of wealth and some income items. It is collected in a harmonised way in 15 euro area countries for a sample of more than 62,000 households. The five countries used in our study account for more than 38,000 of these households. We use macroeconomic data from the FA which are part of the SNA. They provide aggregated macro-level balance sheet data for institutional sectors, including the household sector.

This section is divided into three parts. The first part discusses the conceptual linkages, that is, the linkage between the different assets and liabilities items as they are reported in the HFCS and the FA. To facilitate the discussion, we refer to the items with their exact name labels as they are coded in those two sources. Financial asset items in the FA are coded combining the letter F with a number whereas, in the HFCS they are coded with the letter HD with a number (collected on the household level), PF with a number (collected for each person of the household aged 16 and older) and D for derived items (e.g., aggregated). We use the current national accounts system of the European Union (ESA2010). For the complete list of codes, see System of National Accounts 2008 (SNA2008) and ECB (2012).

The second part focuses on generic differences that have a potential effect on how well the aggregates derived from the survey are able to match the aggregates of FA. Finally, in the third part we derive different adjusted concepts of financial assets that aim to provide a more comparable picture of financial assets than a purely naïve comparison can provide. The purpose of this section is to quantify generic conceptual and statistical differences (see also a similar discussion in Kavonius and Törmälehto 2010 based on ESA95).

2.1. Conceptual Linkages

Although the HFCS uses concepts that are aligned to the FA where possible, the exact definitions sometimes differ to fit the purpose of the questionnaire, as data have to be collected so that households can understand the questions and provide the appropriate information. This might involve asking households about assets or liabilities that do not fit the FA breakdowns, or skipping some items entirely, for concerns that have to do with the interviewing process. This is, for example, the case with currency that is only reported in the FA under item F.21 Currency but is not collected in the HFCS. Asking in a survey about currency at home is generally seen as too sensitive or intrusive.

Table 1 provides an overview of the balance sheet of the FA and the HFCS, only including the items that are relevant for households. The table also indicates items that are not collected in either of the two sources (e.g., currency). Furthermore, Table A1 in Section 6 Appendix shows the linkages on a more detailed financial instrument level. This represents an updated table as shown in Kavonius and Törmälehto (2010), with some refinements of their linkages and changes that came to light through the change from ESA95 to ESA2010.

To compare coverage of both sources, we will define below an "adjusted concept of financial assets." Especially those assets and liabilities that are not covered in either of the two sources have to be first eliminated in defining such a concept to make both sources as comparable as possible. But also assets and liabilities that are hard to compare would have to be excluded to not distort the comparability on an aggregated level. Table A1 in Section 6 Appendix also gives more detail on the financial instruments that we excluded

Table 1. Overview of the balance sheets in the financial accounts and the HFCS.

FA (ESA 2010)	HFCS
Financial assets (+)	
F.21 Currency	N/A
F.22 + F.29 Deposits	HD1110 + HD1210 Deposits
F.3 Debt Securities	HD1420 Bonds and other debt securities
F.4 Loans	HD1710 Money owed to household
F.5 Equity and investment fund shares	HD1510 Shares, publicly traded
	HD1010 Investment in non-self-employed business
	HD0200 Investment in self-employed business[1]
	HD1320x Mutual Funds
F.6 Insurance, pension and standardised guarantee schemes	PF0920 Voluntary pension/whole life insurance schemes PF0700 Occupational Pension Plans[2]
F.7 Financial derivatives and employee stock options	HD1920 Other financial assets
F.8 Other accounts receivable	
N/A	HD1620 Managed Accounts
Liabilities (−)	
F.4 Loans	DL1100 Mortgages and loans
	DL1200 Other, nonmortgage debt (Outstanding debts on credit cards, credit lines and overdraft balances, Noncollateralised loans)
F.8 Other accounts payable	N/A
Financial net worth	
Nonfinancial assets (+)	
N.111 Dwellings	HB0900 Household main residence
N.112 Other buildings/structures	HB28$x + HB2900 Other properties
N.113 Machinery and equipment	N/A
N.13 Valuables	HB4710 Valuables
N/A	HB4400 + HB4600 Vehicles
N.211 Land	N/A (included in entries above)
Net worth	

[1]HD0200 is classified as real wealth in the survey. [2]Usually excluded in the survey definition of financial wealth in the HFCS, but collected in most countries.

from the adjusted concept of financial assets (which we define in Subsection 2.3) and provides a comment for each instrument as regards the comparability. As Kavonius and Törmälehto (2010) have already examined and discussed the linkages between the HFCS and FA, we refrain from a discussion on the linkages on a financial instrument level here.

There are two important differences in the classification to their approach which are worth noting:

First, in the FA the item 'F.51 Equity' consists of the sum of the following three items: 'F.511 Listed shares', 'F.512 Unlisted shares' and 'F.519 Other equity'. Listed shares are equity securities listed on a stock exchange, whereas unlisted shares are equity securities not listed on a stock exchange. Other equity comprises all forms of equity other than listed shares and unlisted shares, e.g., equity in limited liability companies whose owners are partners and not shareholders. For further explanations, see ESA2010, 142–144. The HFCS also collects the value of publicly traded shares (HD1510) that can be linked to F.511 Listed shares'. But contrary to the classification in FA ('Unlisted shares' and 'Other equity'), the classification in the HFCS is based on the household's activity in the enterprise. If the household is self-employed or has an active role in running the business, any unlisted shares or other equity the household would own in the business would be classified in the HFCS as 'HD02000 Investment in self-employed business'. If the household is just invested in the business, for example as a silent partner without having an active role in running the business, and there are no publicly traded shares, then it is classified as a "HD1010 Non-self-employment not publicly traded business". In the HFCS, the value of self-employed businesses is regarded as real wealth, whereas any investments in non-self-employed businesses are regarded as financial assets in the survey classification. To match the categorisation of financial assets in FA we reclassify the value of self-employed businesses to financial assets (other equity).

Second, the SNA2008 introduced new breakdowns for F.6 insurance, pensions and standardised guaranteed schemes, which allows for better linking of the concepts between the HFCS and FA. F.61 Non-life insurance technical reserves are not covered by the HFCS wealth concept. F.62 Life insurance and annuity entitlements correspond with the HFCS item voluntary pensions/whole life insurance schemes. The data in FA is typically based on actuary information on technical reserves reported by insurance corporations. F.63 Pension entitlement corresponds with the HFCS item "current value of all occupational pension plans that have an account" which could be either an amount similar to the present value, or a current (and lower) early liquidation value of the insurance contract (deducting a surrender charge)". However, as the concept in FA does not only cover pensions that have an account balance and as the stock of occupational pensions of households that are already retired is not included in the survey (and in the FA they are), we exclude the pension entitlements in the adjusted concept of financial assets. F.64 Claims of pension funds on pension managers, F.65 Entitlement to non-pension benefits and F.66 Provision for calls under standardised guarantees are not included as it is not considered to be relevant for the comparison.

While we would like to include nonfinancial assets in our analysis, the ESA Transmission Programme requires the transmission of annual data on land only by end-2017. Therefore, this gap in the national accounts data transmission makes it impossible for us to include these in our analysis.

2.2. *Generic Differences*

This section focuses on the generic differences between the HFCS and FA. While the conceptual linkage is important for pointing out differences in definition and for excluding

asset classes that are not comparable, by generic differences we refer to differences that potentially affect all assets and liabilities, though to a different extent. We briefly go through the following differences: (1) population differences; (2) timing; (3) potential measurement errors in the FA; (4) underreporting and item nonresponse in the HFCS; and (5) differences caused by the treatment of sole-proprietors/partnerships and quasi-corporations.

2.2.1. Population

In the comparisons of FA and the HFCS, there are potentially two generic differences with regard to the population: (1) The difference caused by the fact that nonprofit institutions serving households (NPISH) are reported in FA in the same aggregate with households. However, in the euro area countries this is less of an issue as most of the countries transmit the households separately from the NPISH. This is also the case in the countries that are discussed in this article. (2) Differences in the definition of the household sector and the HFCS population. FA have a resident approach, covering all households that plan to stay for at least one year, and irrespective of periods spent abroad of less than one year. In the HFCS, nonresident citizens are not excluded in all countries. In the HFCS, persons living in institutions, for example in prisons or retirement homes, are excluded in most countries; persons with the intention of staying less than six months in the country are also excluded from the target population. Therefore, the household weights, which are designed to represent the target population, do not include these specific excluded groups in most countries. Any comparison has to take this into account and the country totals of the survey or FA have to be adjusted. As an estimate using per capita amounts seems reasonable, with the caveat that this assumes that the excluded groups have the same average wealth as the rest of the population, which may not be the case. For instance, people living in retirement homes may have a per capita wealth that differs from the average.

Table 2 compares the population numbers between FA and the HFCS. The number for FA is based on the last available vintage that corresponds to the reference year of the fieldwork period and is based on the European Commission's ESA95 Transmission Programme population data. Because of the above mentioned excluded groups, the population in the HFCS should generally be lower than the one for the whole population. This is the case for all countries except for Spain. The reason for the "negative" difference is that the Spanish census results have been revised after the first wave results.

Table 2. Comparison of population between FA (ESA95 population data) and HFCS.

Country	Population FA (historical vintage)	Target population HFCS	Difference total	Difference in %
Austria	8,388,130	8,021,945	366,185	4
Germany	81,629,370	81,085,984	543,386	1
Spain	45,456,960	45,632,180	− 175,220	0
Finland	5,336,910	5,271,534	65,376	1
France	64,444,520[1]	62,464,244	1,980,276	3

[1]French overseas territories are included in the FA, whereas the HFCS only includes metropolitan France.

2.2.2. Timing and Frequency

The primary drawbacks of the HFCS are the biennial to triennial frequency and the lag between data collection and data release. Furthermore, the different fieldwork periods may raise concerns about comparability on an aggregated level. The first wave of the HFCS was carried out from 2008/2009 to 2011. For the comparison of the FA with the HFCS, FA data that are closest to the mean of the fieldwork period for each country are used. This is based on annual (year-end) figures as some EU countries do not yet provide quarterly FA backdata for ESA2010, which would better match the fieldwork period of the first wave. Table 3 gives an overview of the different fieldwork periods and the annual end date for FA which is taken for the comparison. The timing can contribute to any observed difference, as the value of assets and liabilities may change between the time the survey was conducted and the period taken for FA.

2.2.3. Potential Measurement Errors in the FA Data

As the FA is based on other statistical sources and the validation of primary statistics, it is possible that errors are inherited from source statistics. Additionally, as the FA is a closed and balanced system, it is possible that some of the household aggregates are adjusted by adding balancing adjustments. In some cases, balance sheet items can even be based on residual estimations. However, in the euro area countries, and in particular in countries that we analyse in this article, the FA balance sheets are mostly based on counterpart information. Although such data might usually be thought of as being relatively accurate, even counterpart information can contain errors. Also, one might not be able to identify the right sector to classify data for all counterpart data (e.g., between S.11 Nonfinancial corporations and S.14 Households). Potential measurement errors in the FA are also discussed in Kavonius and Törmälehto (2010) and Kavonius and Honkkila (2013).

2.2.4. Underreporting and Item Nonresponse in the HFCS

Item nonresponse refers to the problem that for some assets and liabilities the household may not report any value. There are several approaches to alleviate this issue. In the HFCS, the problem of item nonresponse is tackled by multiple imputation, which is the leading method (Rubin 2004). This means that the HFCS, instead of providing one imputed value for each missing one, is providing a set of values drawn from the distribution of values, conditional on the characteristics of the household and the other variables. A full data set

Table 3. Fieldwork period and time periods for comparison.

Country	Fieldwork	Assets and liabilities	FA (annual end)
Austria	Sept. 2010 – May 2011	Time of interview	Q4/2010
Germany	Sept. 2010 – July 2011	Time of interview	Q4/2010
Spain	Nov. 2008 – July 2009	Time of interview	Q4/2008
Finland	Jan. 2010 – May 2010	2009-12-31	Q4/2009
France	Oct. 2009 – Feb. 2010	Time of interview	Q4/2009

Note: Source of fieldwork period, Assets & Liabilities is ECB (2013).

for the main financial instruments without missing values is provided (ECB 2013). This reduces the overall coverage problem between the survey and FA for these items, as the imputed values increase the total amounts of the survey accordingly. One measurement problem that remains apart from item nonresponse is that the household still may not accurately estimate the value of some assets or liabilities, or denies that it possesses the financial instrument. This might also be one reason for discrepancies between the HFCS and FA.

2.2.5. Differences Caused by the Treatment of Sole-Proprietors/Partnerships and Quasi-Corporations

FA distinguishes between producer households (to be classified within the household sector/S.14) and quasi-corporations (to be classified within the nonfinancial corporations sector S.11). This distinction is relevant because it affects the gross wealth of the household sector and the composition of the household balance sheet. In the FA framework it depends whether the business is a separate institutional unit or not: "*Quasi-corporations are unincorporated enterprises that function as if they were corporations. Quasi-corporations are treated as corporations: that is, as separate institutional units from the units to which they belong in recognition of their distinct economic and financial behaviour*." (ESA2010, 422). Unincorporated enterprises are part of the household sector (S.14) and are classified as producer households if they are not considered as a separate institutional unit as described above.

Financial and nonfinancial assets, as well as financial liabilities of these unincorporated enterprises, are spread over the various items of the household balance sheets and it is not possible to distinguish between wealth of the unincorporated enterprise and wealth of the household. In this case, there is no value of net equity recorded in 'F.519 Other Equity'. However, it the economic activity is considered to be a separate unit, any property rights are classified in FA as equity participation held by the household (other equity) and this separate institutional unit is then classified in S.11 or S.12.

The survey definition of self-employed businesses (including sole-proprietorships and partnerships) ideally enables identifying values for the net value of the business separately from other nonbusiness related positions of the household. This conceptual difference implies that for producer households, there is a net value collected in the survey, whereas in FA the assets and liabilities of these producer households are spread over the different instruments. The question is which instruments are affected by this difference and to what extent. Real assets and liabilities may as well be affected as financial assets. To have a measure on the size of this difference for each of the instruments would require separate accounts for sole-proprietorships and partnerships. This might account for part of the difference in the coverage ratios of many instruments, as well as on an aggregated level for each component of net wealth (financial assets, real assets and liabilities). For legal forms other than sole-proprietorships and partnerships (e.g., limited liability companies) the household holds a net equity position in the business both in the FA and in the HFCS.

Table 4 provides an overview of the different types of businesses and how they are recorded in the HFCS and FA. As can be seen, the main comparability issue arises only for those sole-proprietors and partnerships which are not classified as quasi-corporations and hence are recorded in the household sector in FA indistinguishable from the "private part

Table 4. Recording of businesses and inclusion in the different concepts.

Case	Type of business	Net/Gross value		Included/excluded in the HFCS			Comment
		HFCS	FA	Naïve comparison	Adjusted concept 1	Adjusted concept 2	
1	Sole proprietorships and partnerships that are *not* classified as quasi-corporations in FA	Net value	Gross Values: Recorded in the household sector indistinguishable from the "private part of the household". The assets and liabilities of the business part are distributed across the household balance sheet (including financial assets, real assets and liabilities)	Excluded	Included	Excluded	The net value might include real assets
2	Sole proprietorships and partnerships that could be classified as quasi-corporations in FA.	Net value	Net value (other equity)	Excluded	Included	Excluded	In principle comparable but quasi-corporations would be difficult to identify in the HFCS based on the information provided in the survey.
3	Limited liability companies and other incorporated businesses	Net value	Net value (other equity)	Excluded	Included	Included	

of the household" (case 1). For these, there is a net value for the business provided in the HFCS, whereas in the FA the assets and liabilities of the business are spread across the balance sheet of the household sector including real assets and liabilities. Hence, for this part of the sole proprietors and partnerships it is not known if the net value of the business provided in the HFCS should be allocated to financial assets, real assets or liabilities in FA. For quasi-corporations (case 2), there is a net value provided in the HFCS and also a net value recorded in the FA. The same applies to the other incorporated businesses: there is, in principle, no difference in the recording, as there is a net value provided in both the HFCS and the FA, although differences in the valuation might still occur.

Furthermore, Table 4 provides an overview of whether the described cases are included or excluded in the HFCS in each of the concepts described in the next section. For the other instruments, Table A1 in Section 6 provides an overview of which instruments are excluded from both sources in the adjusted concepts.

2.3. Adjusted Concepts of Financial Assets

The aim of this section is to derive two adjusted concepts of financial assets. The intention to go from a naïve comparison to an adjusted concept is done by basing the comparison on those financial instruments which are included in both sources and are conceptually comparable. The adjusted concepts allow providing a more reliable indication of those financial assets that are covered in both the HFCS and the FA. We define the 'coverage ratio' as measuring the per capita amount of financial assets covered by the survey, for example a value of 98% would imply that the per capita amount of the HFCS is only 2% below the per capita amount in FA.

$$Coverage\ Ratio = \frac{DA2100}{AF} \tag{1}$$

where AF refers to the total financial assets in the FA and DA2100 refers to the total financial assets in the HFCS.

2.3.1. Naïve Comparison

The naïve comparison takes the concepts of financial assets as they are in the HFCS and in the FA. This serves a benchmark, but this concept also includes noncomparable instruments (e.g., F.21 Currency, which is not covered by the HFCS) and uses different classifications (e.g., the value of self-employed) that distort the picture of the actual coverage ratios. The HFCS concept of financial assets does not include the value of self-employed businesses, as well as the value of occupational pension plans that are accordingly also not included in the naïve comparison. Therefore, it is not surprising that the naïve comparison shows relatively low coverage ratios of 34% to 43% for financial assets (results are presented in Table 5).

FA:

$$AF_{Naïve} = F.21 + F.22 + F.29 + F.3 + F.4 + F.5 + F.6 + F.7 + F.8 \tag{2}$$

Table 5. Coverage ratios of financial assets for the household sector (S.14) – Naïve comparison vs. adjusted concepts.

Country	Coverage ratio (%)			Share of total financial assets in the FA covered in the adjusted concepts (same in both concepts) (%)
	Naïve comparison	Adjusted concept 1	Adjusted concept 2	
Austria	35	98	46	87
Germany	43	86	67	77
Spain	34	75	59	82
Finland	37	55	45	83
France	38	59	51	90

Notes: The coverage ratio of the different concepts is reported. The naïve comparison includes all assets as given in the two sources, without taking into account the conceptual comparability. For the adjusted Concepts 1 and 2, we make adjustments to increase the conceptual comparability. The share of total FA shows the assets covered in the adjusted Concepts 1 and 2 as a percentage of total financial assets in the financial accounts (same for both concepts). *Sources*: HFCS and Financial Accounts.

HFCS:

$$DA2100_{Naïve} = HD1110 + HD1210 + HD1320x + HD1420 + HD1010 + HD1510 + HD1620 + HD1710 + HD1920 + DA2109 \quad (3)$$

2.3.2. Adjusted Concept 1

For the adjusted Concept 1 we include on the survey side the value of self-employed businesses (DA1140 (which is the sum of (HD080x) + HD0900)) in the comparison (reclassification from real assets to financial assets) and we exclude the amount owed to the household (HD1710), as well as the other financial assets (HD1920). In the FA, we exclude F.21 Currency, F.4 Loans (Assets), F.7 Financial derivatives, and F.8 Other accounts receivable. For pensions, we only include F.62 Life insurance and annuity entitlements and exclude the other subcategories (F.61, F.63-F.66) as these are not comparable to the survey (see discussion above). As can be seen in Table 5, going from a naïve comparison to the adjusted Concept 1 significantly increases the coverage ratio for financial assets (to 55% in Finland and even 98% in Austria). Putting these numbers in perspective, it is worth noting that in their comparison of the Flow of funds Accounts (FFA) and the Survey of Consumer Finances (SCF) in the US, Henriques and Hsu (2014) conclude that the net worth of the SCF in comparable terms is above the net worth of the FFA. More recently, Bricker et al. (2016b) show that much of the wealth gap between the SCF net wealth and FA wealth seems to be for assets where market prices are not easily observed. For example Bricker et al. (2016b) show that in 2013, SCF housing was 36% above the FA estimate, but SCF nonhousing assets were only 6% above the FA.

The adjusted Concept 1, as it includes all self-employed businesses, most likely overstates the coverage ratio, as the value for sole proprietors and partnerships may also include real assets (see discussion above about the delineation between sole-proprietors and quasi-corporations).

FA:

$$AF_{adj1} = F_{Naïve} - F.21 - F.4 - F.61 - F.63 - F.64 - F.65 - F.66 - F.7 - F.8 \quad (4)$$

HFCS:

$$DA2100_{adj1} = DA2100_{Naïve} - HD1710 - HD1920 + DA1140 \quad (5)$$

2.3.3. Adjusted Concept 2

In the HFCS, the value for self-employed businesses can be broken down by legal status (see Table A1 in the Section 6). Therefore, the distinction between sole-proprietorships, partnerships and other incorporated businesses is possible. While the adjusted Concept 1 includes the net values of all legal forms of self-employed businesses in the HFCS (including sole proprietors and partnerships), the adjusted Concept 2 excludes sole proprietors and partnerships from the value of self-employed businesses in the survey. For FA, we keep the corresponding instrument F.5 Equity the same in both concepts. The intention of this adjusted concept is that it serves as a lower benchmark, as it only comprises the net value of those legal forms in the survey that are recorded in the nonfinancial corporations' sector in the FA and, consequently, the household only holds a net equity position in the business (other equity). Thus, for the legal forms included in this concept both in the FA and in the HFCS, the household holds a net equity position.

As can be seen in Table 5, the coverage ratios for the adjusted Concept 2 are higher compared to the naïve comparison, but significantly lower compared to the adjusted Concept 1, where all legal forms of self-employed businesses are included. Certainly, adjusted Concept 2 underestimates the coverage ratios, as it excludes all financial assets of sole-proprietorships and partnerships from the survey.

To further improve the comparability between the HFCS and the FA, the following information would be needed: first, an estimate of sole proprietorships and partnerships included in the HFCS that are classified as quasi-corporations in the FA (case 2 in Table 4). Second, for the sole-proprietors and partnerships that are recorded in the household sector, one would need the breakdown to financial assets, real assets and liabilities (case 1 in Table 4).

FA:

$$AF_{adj2} = F_{Naïve} = F.21 - F.4 - F.61 - F.63 - F.64 - F.65 - F.66 - F.7 - F.8 \quad (6)$$

HFCS:

$$DA2100_{adj2} = DA2100_{Naïve} - HD1710 - HD1920 + DA1140 - DA1140_{Sole\ proprietorships/independent\ professionals+partnerships} \quad (7)$$

3. The Wealth Distribution and Methodology to Estimate the Tail

In this section, we first discuss the general problem of wealth surveys, that is, the fact that top tail observations are missing, which is often caused by differential unit nonresponse. We also discuss which oversampling strategies are used by countries to mitigate this issue in the HFCS. In the second parts, we explain the methodology to estimate the top tail of the

wealth distribution by a Pareto distribution. Our approach and discussion builds on Vermeulen (2018). The third part discusses the Forbes list and its consistency with the statistical data. These data are used for the estimations of the Pareto tail.

3.1. Oversampling Wealth Distribution and Differential Unit Nonresponse in the HFCS

In general, the bias in the HFCS caused by unit nonresponse is reduced by weight adjustments (Pérez-Duarte et al. 2010). But as the wealth distribution is often skewed, unit nonresponse of the wealthiest households, or the fact that the extremely wealthy households are rarely included in the survey sample, is still usually a problem. Income and wealth concentrations are likely to be underestimated using survey data, as there is a high concentration of wealth in the top quintile and the response rates of this quintile, in particular, is usually lower. For the top tail of the wealth distribution, there is some evidence on how response rates correlate with the amount of wealth owed by a household. Based on the Survey of Consumer Finance from the United States, Kennickell and Woodburn (1999) have documented the following response rates based on different strata (differential unit nonresponse): 34% for USD 1 million to USD 2.5 million and 14% for USD 100 million to USD 250 million. For the stratum that likely includes the wealthiest households, Kennickell (2008) observes an overall response rate of 10%. Bricker et al. (2016b) report response rates for more recent SCF waves in the wealthiest SCF stratum of around 12%, around 25% in the second stratum, rising to around 50% in the last two least-wealthy strata. This is still lower than the response rate of around 70% in the SCF area probability sample. However, Bricker et al. (2016a) nicely demonstrate that even though response rates are low at the top of the wealth distribution, the survey participants are observationally equivalent to the nonrespondents. This demonstrates the usefulness and effectiveness of oversampling.

For the HFCS, the amount of wealth owned by the top tail varies from country to country and available evidence suggests that the response rates declined to a different extent in different countries. For the 2011 wave, the Spanish survey of household finances documented the following response rates by wealth strata: Stratum 5 (0.9 to 2 million) 31%, Stratum 6 (2 to 6 million) 26%, Stratum 7 (6 to 25 million) 21% and Stratum 8 (wealth above EUR 25 million) 21%. The survey also has a panel component, for which the response rate drops from 74% to 62% for these wealth strata (Bover et al. 2014). On the other hand, in Finland – although response rates varied across different strata, age groups, regions and education level, nonresponse rates did not increase along the level of taxable wealth for the Finnish Household Wealth Survey of 2004 (Pérez-Duarte et al. 2010).

Some countries have oversampled wealthy households in the HFCS to increase the precision at the top. Table 6 gives an overview of the oversampling strategy for the countries included in our analysis. Germany used an oversampling strategy based on geographical areas that resulted in a less effective oversampling than in France and Spain, which used net wealth or taxable wealth. One should expect that oversampling increases the precision of the aggregated survey values and therefore make them potentially closer to the FA for a single survey.

Even with oversampling, it remains uncertain how much of the wealth of the wealthiest households is actually covered by the survey. This in turn, is one reason for part of the gap

Table 6. Oversampling in the first wave of the HFCS by country.

Country	Oversampling wealthy households	Basis for oversampling	Effective oversampling rate of the top 5%
Austria	No	N/A	4
Germany	Yes	Geographical areas	148
Spain	Yes	Taxable wealth	314
Finland	Yes	High-income employees, self-employed and farmers	85
France	Yes	Net wealth	208

Notes: The source is HFCS. Effective oversampling rate of the top 5%, (S95 – 0.05)/0.05, where S95 is the share of sample households in the wealthiest 5%. Wealthiest households are defined as having higher net wealth than 95% of all households, calculated from weighted data (ECB 2013).

between the amounts of FA and aggregated amounts from the survey. The methodology presented in the next section addresses exactly this issue. The idea is to replace the observations above a certain threshold of net wealth per household by an estimated Pareto distribution and see which impact this has on the coverage ratio of the HFCS in comparison to FA. In terms of the coverage ratio, capturing the value of assets from these wealthiest households might be even more relevant for specific instruments, as there are particular financial assets that are largely owned by a small fraction of the wealthier households. Here, we concentrate on net wealth figures, as well as on the adjusted concept of financial assets and leave the breakdown on particular instruments for future research (see Chakraborty and Waltl 2018). The methodology used to estimate the Pareto tail is the same approach as in Vermeulen (2018). Therefore, we keep the explanation here short.

3.2. Methodology

Wealth is heavily skewed at the top and the literature has reached a consensus that the top of the wealth distribution is well approximated by a Pareto distribution (Davies and Shorrocks 1999). The Pareto distribution has two parameters, the tail exponent α and the threshold parameter T. The distribution is given by the following complementary cumulative distribution function (ccdf):

$$P(W > w) = \left(\frac{T}{w}\right)\alpha \tag{8}$$

The Pareto distribution is defined on the interval [T,∞) and $\alpha > 0$. The threshold T is the lower bound of the distribution. Estimating a Pareto distribution on a simple random sample is fairly straightforward. The maximum likelihood estimator of α from a random sample of n observations drawn from a Pareto distribution with a given threshold T is given by:

$$\alpha_{ml} = \left[\sum_i \frac{1}{n} \ln \frac{w_i}{T}\right]^{(-1)} \tag{9}$$

Alternatively, the tail exponent has been estimated in the literature using linear regression on ranked data. Let i be the rank of the observation (with rank 1 being the highest observation). The Pareto tail exponent α can be estimated by:

$$\ln(i - 0.5) = C - \alpha \cdot \ln(w_i) \tag{10}$$

Where the "subtract 0.5 from the rank" is suggested in Gabaix and Ibragimov (2011).

However, wealth survey data generally does not consist of a simple random sample. In particular, sample observations have weights. Vermeulen (2018) shows that taking into account the weights can be done in the regression method above, using the ranked n highest observations:

$$\ln(i - 0.5) = \frac{N_{fi}}{\bar{N}} C - \alpha \cdot \ln(w_i) \tag{11}$$

where N_{fi} is the average weight of the highest i sample points and $\bar{N}$ is the average weight of all n highest sample points. This regression method can be used in two ways. First, estimate α using only the survey data (i.e., the highest n observations). Alternatively, these observations can be pooled with data of rich lists that contain datapoints that are higher than the highest observation in the survey (this joint data set is then ordered first). Using this regression method works particularly well when combining the survey data with such extraneous data points.

A particular problem is the choice of the threshold T. There is no clear-cut way in finding a "correct" threshold. However, the Pareto distribution has the interesting property that a distribution with tail exponent α and threshold T, when restricted above $T* > T$ remains a Pareto with the same tail exponent. Therefore, it seems prudent with survey data to use a high threshold. This way, lower observations that are not Pareto distributed are avoided. However, there is a trade-off: a higher threshold $T*$ implies using less data to estimate α. It is probably best to estimate α using different thresholds of the data and check for sensitivity.

After estimating the α for a given threshold T, the n observations can be replaced by the estimated Pareto distribution. The mean of a Pareto distribution is given by $\frac{\alpha}{\alpha-1} T$, so that we can say that the total wealth in the Pareto tail is given by $n\bar{N}\frac{\alpha}{\alpha-1} T$, where $n\bar{N}$ is the total sum of weights of the highest n observations in the survey sample.

We use the thresholds EUR 500,000, EUR 1 million, and EUR 2 million to estimate α and we use the same thresholds to replace the survey observations by the estimated Pareto tail. The Pareto distribution is estimated using the above described methods: (1) the pseudo maximum likelihood. Specifically, we use the pseudo-maximum likelihood estimator which has the same form as the maximum likelihood estimator, but takes into account the weights of the sample observations in the survey (see Vermeulen 2018); (2) the regression method excluding data from the Forbes; and (3) the regression method including data from the Forbes.

3.3. *Forbes Data*

The wealth concept of the Forbes list does not strictly follow any defined concept and therefore, it should be interpreted as a proxy. The wealth concept typically covers the net

wealth and thus, the split between assets and liabilities is not available. Four conceptual issues related to the use of these estimates in statistical estimations can be identified. First, the estimations are based either on interviews of billionaires themselves or their handlers, employees, rivals, or others. This implies that it is impossible to cover all the asset types or to have types of market valuation that are similar to FA or household surveys. On the methodology used by Forbes, Dolan (2016) states that "*not that we pretend to know what is listed on everyone's private balance sheet, though some folks do provide that information. We do attempt to vet these numbers with all billionaires. Some cooperate, others don't.*" Almost all the families on the Forbes list from the countries analysed in this article have earned their money in businesses and therefore, it can be assumed that the majority of their net wealth is in equity. For the Forbes list, the privately-owned businesses have been valued by coupling estimates of revenues or profits with prevailing price-to-revenues or price-to-earnings ratios for similar companies (Dolan 2016). This method can be considered similar to the methods used in the valuation of the unlisted equity in the FA.

Second, the wealth concept does not cover all asset types, as these are partly based on external estimations. Additionally, the wealth concept also covers items that are defined as durable goods in the NA (such as yachts). Third, sometimes the fortune is distributed to the different family members and sometimes it is not, and a large number of family members is aggregated (Dolan 2016). The starting point in statistics and in particular in the HFCS is that the applied unit is the household. In the case of the Forbes list, it is very possible that the applied family concept covers several households or reversely, one person, for example, the head of the household.

Fourth, the Forbes list covers families by nationality and it does not correspond with the residence concept applied in the HFCN and the SNA. The families living outside of the country of their citizenship should not be included in the HFCN population, but they are included in the Forbes list. A brief analysis proved that the majority of these families are actually resident in the countries of their citizenship. For instance, in the case of Finland, all six persons who are on the list are also residents in Finland. In larger countries, where the number of billionaires is also higher, there are some families that live outside the country of their citizenship. In future work, allocating these types of families to their resident countries can be considered. Even though there are these drawbacks in using the Forbes list, the data are one of the best proxies for the very top tail of the wealthiest households (alternatives being national rich lists).

4. Results

4.1. Estimates of the Pareto Compared with the HFCS

We estimate the Pareto tail exponent using the three methods described above, for the three thresholds. The results for the Pareto tail exponent (alpha) are provided in Table 7. The Pareto tail index estimates coincide with those found by Vermeulen (2018). In general, a lower α implies higher tail net wealth and higher total net wealth. As described earlier, we replace the tail net wealth of the survey observations above each of the

Table 7. Pareto tail index (alpha).

	Pseudo max.likelihood			Regression method excl. Forbes			Regression method incl. Forbes		
Country	≥2M	≥1M	≥500T	≥2M	≥1M	≥500T	≥2M	≥1M	≥500T
Austria	1.67 (0.42)	1.42 (0.30)	1.34 (0.16)	1.87 (0.72)	1.65 (0.45)	1.44 (0.26)	1.47 (0.06)	1.47 (0.05)	1.43 (0.08)
Germany	1.41 (0.26)	1.43 (0.17)	1.61 (0.10)	1.87 (0.35)	1.64 (0.23)	1.54 (0.13)	1.38 (0.04)	1.39 (0.02)	1.40 (0.01)
Spain	1.71 (0.27)	2.05 (0.18)	1.85 (0.08)	1.67 (0.13)	1.76 (0.11)	1.87 (0.08)	1.59 (0.05)	1.69 (0.05)	1.80 (0.05)
Finland	2.01 (0.23)	2.47 (0.18)	2.26 (0.06)	1.94 (0.57)	2.13 (0.23)	2.27 (0.10)	1.60 (0.14)	1.88 (0.13)	2.16 (0.08)
France	1.65 (0.09)	1.84 (0.08)	1.75 (0.04)	1.67 (0.13)	1.78 (0.08)	1.83 (0.06)	1.50 (0.02)	1.63 (0.03)	1.73 (0.03)

Notes: The table shows the results of three different methods used to estimate the Pareto tail index (alpha). For each of the three methods, we vary the threshold used for the estimation of the alpha. Mean over the results computed in each of the five implicates and standard errors are shown in parentheses. Sources: HFCS and Forbes.

Table 8. Weights below and above threshold.

Country	≥2M		≥1M		≥500T	
	Below	Above	Below	Above	Below	Above
Austria	0.981	0.019	0.954	0.046	0.887	0.113
Germany	0.991	0.009	0.974	0.026	0.918	0.082
Spain	0.992	0.008	0.964	0.036	0.865	0.135
Finland	0.997	0.003	0.986	0.014	0.937	0.063
France	0.992	0.008	0.970	0.030	0.896	0.104

Notes: The table shows the weights allocated in the HFCS above and below three different thresholds for the given countries. The thresholds refer to net wealth. The sum of the weights corresponds to the size of the target population (see Table 2). *Source*: HFCS.

thresholds by the estimated net wealth from the Pareto distribution. Thus, we assume that the weights in the HFCS allocated to those households having net wealth above these thresholds are correct.

Table 8 gives an overview of the weights in % of the population. Obviously, increasing the threshold decreases the weights allocated to these households. Nevertheless, the weights allocated to households above each of the threshold varies from country to country, for example, the weight of households in Austria with a net wealth above EUR 2 million is 1.9% compared to Finland with only 0.3%.

Table 9 shows the net wealth below and above the thresholds as measured in the HFCS.

Table 10 to Table 12 provide the estimates of tail net wealth using the different methods to estimate the tail. They also provide a comparison in terms of the HFCS tail for each estimate (Pareto tail divided by the HFCS tail). Furthermore, the tables provide an estimate in terms of actual net wealth of the HFCS when the tail is replaced by the Pareto estimate (estimated net wealth when tail is replaced by the Pareto divided by the actual net wealth of the HFCS).

Table 10 shows the tail net wealth using the *pseudo maximum likelihood method* without the Forbes list. The tail does not significantly increase for those countries that used an effective oversampling strategy (Spain and France). However, especially in Austria and Germany, with less effective oversampling strategies, the estimated Pareto tail increase

Table 9. Net wealth below and above threshold HFCS (EUR billions).

Country	≥2M		≥1M		≥500T		Total
	Below	Above	Below	Above	Below	Above	
Austria	673	327	528	472	357	643	1,000
Germany	5,907	1,836	4,945	2,798	3,489	4,254	7,743
Spain	4,273	685	3,637	1,321	2,475	2,483	4,958
Finland	384	25	349	60	267	142	409
France	5,466	1,036	4,620	1,883	3,200	3,303	6,503

Notes: The table shows the net wealth, aggregated over households below and above the threshold as it is given in the HFCS. The thresholds refer to net wealth. The total shows the aggregated net wealth in the HFCS for each country. *Source*: HFCS.

Table 10. Estimated net wealth above threshold (tail wealth) using Pseudo max. likelihood.

	≥2M			≥1M			≥500T		
Country	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth
Austria	354	108	103	590	125	112	842	131	120
Germany	2,536	138	109	3,496	125	109	4,304	101	101
Spain	672	98	100	1,213	92	98	2,503	101	100
Finland	26	106	100	58	96	99	142	100	100
France	1,064	103	100	1,820	97	99	3,374	102	101

Table 11. Estimated net wealth above threshold (tail wealth) using regression method excluding Forbes.

Country	≥2M			≥1M			≥500T		
	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail net wealth	Estimated net wealth in % of HFCS net wealth
Austria	305	93	98	443	94	97	699	109	106
Germany	1,585	86	97	2,694	96	99	4,651	109	105
Spain	696	102	100	1,438	109	102	2,472	100	100
Finland	27	110	101	65	108	101	142	100	100
France	1,045	101	100	1,896	101	100	3,188	97	98

Table 12. Estimated net wealth above threshold (tail wealth) using regression method including Forbes.

Country	≥2M			≥1M			≥500T		
	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail wealth	Estimated net wealth in % of HFCS net wealth	Tail net wealth (bn EUR)	Tail net wealth in % of HFCS tail net wealth	Estimated net wealth in % of HFCS net wealth
Austria	444	136	112	546	116	107	710	110	107
Germany	2,678	146	111	3,747	134	112	5,708	134	119
Spain	752	110	101	1,521	115	104	2,587	104	102
Finland	35	141	103	74	123	103	148	104	101
France	1,258	121	103	2,149	114	104	3,427	104	102

Notes to Tables 10 to 12: The tail net wealth shows the net wealth estimated by each of the methods using the three specified thresholds. The tail net wealth in % of HFCS tail divides the estimated tail net wealth above the specified threshold by the tail net wealth as it is measured in the HFCS. The estimated net wealth in % of HFCS net wealth takes the estimated net wealth when the tail is estimated and aggregated together with the net wealth below the threshold from the survey and divides the sum by the aggregated net wealth as it is measured in the HFCS.

the tail compared to the HFCS, as well as total net wealth. The total effect on net wealth is lower compared to the effect on the tail, as the weight of the households with net wealth above each of the thresholds is taken into account.

Table 11 shows the tail net wealth using the *regression method excluding the Forbes list.* Similarly, not much is added to the HFCS tail using either of the thresholds for the countries with an effective oversampling strategy. Generally, the estimates get more imprecise the higher the threshold is, as fewer sample observations from the survey can be used for the analysis. This seems to be especially prevalent for the countries using a less effective oversampling strategy. The lower estimated tail for Austria and Germany with a threshold of EUR 2 million is most likely based on this fact, and the results need to be interpreted with caution. However, using a lower threshold generally brings the risk of including observations in the estimate that may not be Pareto distributed. Finally, Table 11 shows the results *including the Forbes data into the regression method.* This yields the highest estimates for the tail, as well as for net wealth, in line with the results from Vermeulen (2018). It even adds net wealth for the countries that used an effective oversampling strategy, although to a minor extent compared to the countries with a less effective oversampling strategy.

In the next section, we analyse how replacing the tail by the Pareto distribution changes the coverage ratios of the adjusted concepts discussed above. We limit the analysis to the last estimation method, including the Forbes list and the adjusted concept 1. For the calculations, we take the mean over the results computed in each of the five implicates provided in the HFCS.

4.2. Comparison with the Adjusted Concept of Financial Assets

In the previous section the tail of the wealth distribution was estimated by taking net wealth as the underlying concept. But so far, we have not broken down the net wealth into its components – financial assets, real assets and liabilities. To make these estimates comparable to the adjusted concept of financial assets discussed in Subsection 2.3, we need to allocate the estimated tail net wealth to financial assets, real assets and liabilities.

To obtain a first estimate, we use the HFCS to calculate the aggregate shares of financial assets, real assets and liabilities for those households that have net wealth above each of the thresholds. Using those shares, we can allocate the Pareto tail net wealth. To give an indication how this allocation changes with net wealth, Table 12 shows the shares above each of the thresholds constructed using the HFCS. The share of financial assets increases, while the share of real assets decreases with a higher threshold of net wealth for all countries included in the study. For these households, liabilities play a minor role (1% – 6%). For the breakdowns provided in Table 13, we have already reclassified self-employed businesses to financial assets.

In Table 14, we show a finer breakdown of financial assets for the households in the survey with net wealth above the threshold of EUR 2 million. One sees that the large part of net wealth for these households consists of the value of self-employed businesses (representing 28% of net wealth in France versus 51% in Austria).

We suspect that the share of financial assets and, in particular, equity increases further for the wealthier households that are not included in the survey. We base this conjecture on

Table 13. Total share of assets and liabilities for households above different thresholds (in % of net wealth).

Country	≥ 2M			≥ 1M			≥ 500T		
	Financial assets	Real assets	Liabilities	Financial assets	Real assets	Liabilities	Financial assets	Real assets	Liabilities
Austria	59	42	1	57	44	1	50	52	2
Germany	57	47	3	49	56	5	43	63	6
Spain	49	53	2	39	65	3	31	74	4
Finland	67	39	6	46	59	6	31	76	6
France	57	45	3	48	55	3	39	65	5

Notes: The percentages show the total share of assets and liabilities for those households having net wealth above each threshold. In this breakdown, the value of self-employed businesses has already been classified in the financial assets. *Source*: HFCS.

Table 14. Total share of financial assets for households with net wealth above 2M (in % of net wealth).

	Austria	Germany	Spain	Finland	France
DA2101 Deposits	3	2	5	3	2
DA2102 Mutual Funds	2	2	2	7	2
DA2103 Bonds	2	2	0	1	1
DA1140 Value of Self-Employment Businesses	51	46	33	33	28
DA2104 Value of Non Self-Employment Private Business	0	0	5	0	4
DA2105 Shares, Publicly Traded	0	2	2	22	6
DA2106 Managed Accounts	0	0	0	0	0
DA2107 Money Owed To Households	1	0	1	0	0
DA2108 Other Assets	0	1	0	0	1
DA2109 Voluntary Pension/ Whole Life Insurance	0	2	1	1	13
Total Financial Assets	59	57	49	67	57

Notes: The table shows, in percentages, the asset allocation for households with net wealth above EUR 2 million for financial assets (% in terms of net wealth). *Source*: HFCS.

the fact that it is often owners of large businesses that can be found on rich lists. After allocating net wealth to an instrument level, we apply the same procedure to derive the adjusted Concept 1 – reallocate self-employed businesses to financial assets and, again, exclude the instruments from the adjusted concept that are not comparable or hardly comparable. The effect of this procedure on the coverage ratios is shown in Table 15. The table also shows the change in the coverage ratio compared to Table 5. This can be interpreted as the change in the coverage ratio that is based on replacing the tail by the Pareto estimate.

Table 15. Coverage ratio of adjusted Concept 1 (financial assets) if tail wealth is replaced using regression method including Forbes.

Country	Coverage ratio (%) adjusted concept 1 (≥2M)	Increase in %	Coverage ratio (%) adjusted concept 1 (≥1M)	Increase in %	Coverage ratio (%) adjusted concept 1 (≥500T)	Increase in %
Austria	110	(+12)	105	(+7)	103	(+5)
Germany	100	(+14)	100	(+14)	104	(+18)
Spain	78	(+3)	81	(+6)	78	(+3)
Finland	59	(+4)	59	(+4)	56	(+1)
France	63	(+4)	63	(+4)	60	(+1)

Notes: The brackets show the change in the coverage ratio to the adjusted Concept 1 for the household sector (S.14) when the tail is replaced with the Pareto estimate. *Sources*: HFCS, FA and Forbes.

We only apply this procedure using the regression method, including the Forbes and the adjusted Concept 1. We take the aggregated portfolio structure above each threshold of the households included in the survey. Thus, a lower threshold also implies a lower percentage of net wealth allocated to financial assets, as can be seen in Table 13. The intention is to point out one further measurement problem that arises when breaking down the estimated tail net wealth to financial assets and real assets. As the net wealth of the Forbes almost always originates from listed or unlisted corporations, most likely the large bulk of their net wealth is invested in equity. So, the estimates gained here can only be understood as an indication of the portfolio allocation of the top tail, but most likely the net wealth estimated by the Pareto could be allocated even more to financial assets/equity, reducing the gap for financial assets even further.

As can be seen, adding the estimated tail increases the coverage ratio for all countries, but to a larger extent for countries with a less effective oversampling strategy. For Spain, Finland and France, the increase created by adding the Pareto tail including the Forbes is not sufficient to reduce the gap to FA. For Austria and Germany, applying a threshold of EUR 2 million increased the coverage ratio significantly. To see why this is the case: first, the two countries have lower estimated alphas (for the regression method including Forbes), hence a larger estimated tail. Second, the weights allocated to households above the EUR 2 million threshold are highest in Austria and Germany. Third, the share of financial assets in Austria is relatively high. Apart from using different estimated alphas and weights, using a lower threshold here also means using a lower portfolio share for financial assets. The share of financial assets for households with a net wealth above EUR 500,000 most likely underestimates the share of financial assets of the Pareto tail and thus, also, underestimates the coverage ratios. The adjusted Concept 1 seems to work particular well for Austria and Germany, but one has to keep in mind that two opposing influences still have an impact which have not been estimated here. On the one hand, the value of real assets of sole-proprietors may be included in the adjusted Concept 1 in financial assets. Excluding these real assets would lead to a lower coverage ratio for financial assets. On the other hand, taking a higher portfolio share of financial assets would lead to an even higher coverage ratio for financial assets. This higher portfolio share can be assumed from the discussion on the Forbes and when taking into account the development of the share in financial assets when increasing the threshold.

5. Conclusion

Using data from the HFCS and the FA, we have made a thorough comparison between both sources for financial assets for Austria, Germany, France, Spain and Finland. We have briefly reviewed the linkages between both sources on an instrument level. Furthermore, we have pointed out and partly estimated basic statistical differences between both sources that have a potential effect on the comparability between both sources.

By developing an adjusted concept of financial assets, we have shown that a large part of the gap in comparison to a naïve comparison can already be explained by conceptual differences and by a reclassification of self-employed businesses from real assets to financial assets aligning the concepts of financial and real assets across both sources.

Identifying comparable items is essential for being able to actually calculate more reliable coverage ratios.

One challenge in deriving adjusted concepts for financial assets is the treatment of self-employed businesses. Here, the issue is which part of sole-proprietors and partnerships included in the survey are assigned in FA to the household sector and which ones are classified as quasi-corporations, and hence are recorded in the nonfinancial corporations' sector. In the latter case, the household only holds a net equity position in the business (other equity). On the other hand, if the sole-proprietors and partnerships are recorded in the household sector, the assets and liabilities may be spread over the balance sheet of the household sector and the net value recorded in the survey may very well include real assets and liabilities. Even though this does not have an effect on the coverage ratios in terms of net wealth, it has a significant effect on the coverage ratio of financial assets on an aggregated level, as well as on an instrument level.

Focusing on the wealthiest households, we have used the estimation procedure from Vermeulen (2018) and replaced those observations in the survey (households) above three different thresholds (EUR 500,000, EUR 1 million, and EUR 2 million) by an estimated Pareto tail. Thus, we allocate the same weights to the estimated tail as are allocated to households above these thresholds in the HFCS. Using the estimates from the Pareto, we have shown the effect on the tail itself and the effect on net wealth. For the countries that already use an effective oversampling strategy, the estimates without the Forbes list do not seem to add much to net wealth and to the coverage ratio. For countries with a less effective/no oversampling strategy, the Pareto estimates seem to increase the tail, net wealth and eventually the coverage ratio. This is one of the main contributions of this article: we analyse how the coverage ratios for comparable financial assets (adjusted concept) change when the top tail is replaced by a Pareto distribution including the Forbes list and which factors are of importance for the final results.

It seems that for countries with an effective oversampling strategy, the increase in the coverage ratio is lower than for countries with a less effective oversampling strategy. Apart from oversampling, three factors are relevant for the final results: first, the estimated alpha is crucial, as *a lower estimated alpha leads to a larger estimated tail*. Second, *the weight allocated to households with wealth above the thresholds is different* from country to country and hence, leads to a different effect on net wealth. Third, *the portfolio allocation to financial assets is relevant when net wealth is broken down to its components (financial assets, real assets and liabilities).* Households with higher net wealth seem to be more invested in financial assets. The analysis shows that it is reasonable to assume that the largest part of financial assets of the wealthiest households is equity. This matters for the estimated coverage ratios for financial assets, as a higher portfolio share in financial assets implies that a larger part of the estimated tail wealth is allocated to financial assets.

In the future, we need to continue to work on adjusting the concept of net wealth including real assets and liabilities. For the estimation of the coverage ratio of the different components, it would be valuable to have an estimate on the share of financial assets, real assets and liabilities held by sole-proprietors and partnerships, as this would give an estimate of how much the adjusted concept for financial assets (adjusted Concept 1) overestimates the coverage ratio.

The analysis shows that the threshold for estimating the alpha might be of importance, as the outcome of the Pareto index might be quite different when taking different thresholds. Generally, there is a trade-off, as increasing the threshold decreases the number of households on which the estimates are based. However, taking a lower threshold brings the risk of including observations (households) that are not Pareto distributed. The threshold is of equal importance for taking the portfolio shares of net wealth allocated to financial assets, real assets and liabilities, as this has an impact on the coverage ratio for each component of net wealth. In the analysis of this article, we have kept the thresholds for estimating the alpha and the portfolio shares the same. A sensitivity analysis varying the thresholds and the portfolio shares, for example estimating the alpha based on the EUR 500,000 threshold but varying the share of financial assets held by these households, would be one way to further analyse the effect on the coverage ratios. Although the regression method including the Forbes shows, on average, lower alphas and hence a bigger tail, the coverage ratios crucially depend on the weight allocated to the tail in the survey. This, in turn, is based on the sampling procedure applied by each country. Thus, varying the weight and observing the effect on the coverage ratio would be worth examining, as the weight differs quite a bit between the countries.

Finally, returning back to our initial question stated in the title 'Is the Top Tail of the Wealth Distribution the Missing Link between the Household Finance and Consumption Survey and National Accounts?' The answer is a qualified "yes but partially". We have shown that the estimated Pareto tail might explain part of the coverage ratio for financial assets, but to a lesser extent than we initially expected. For the countries that have a less effective oversampling strategy, a larger part of the gap to FA seems to be explained by the estimated top tail. But apart from the applied oversampling strategy, the change in the coverage ratio depends on the distribution of wealth in each country (leading to different alphas), the weight allocated to households in the top of the distribution, and the portfolio allocation of the wealthy households. Finding the 'correct' estimates for each measurement problem is a difficult task. The question remains, for some countries in our analysis, why the coverage ratios using the adjusted concepts are still relatively low, and further explanations have to be found. One such explanation is underreporting. We leave this for future research.

6. Appendix

Table A1. Correspondence table: financial wealth in HFCS and FA.

ESA 2010 code	FA/Instrument name	HFCS variable code(s)	HFCS variable	Adjusted concept	Remarks
	Assets				
F.2	**Currency and deposits**				
F.21	Currency	N/A	N/A	Excluded	FA: holdings by households included but estimated due to the lack of direct sources. HFCS: Not collected.
F.22 F.29	Transferable deposits Other deposits	HD1110 HD1210	Sight accounts Savings accounts	Included Included	Specific conceptual differences exist but are unlikely to be significant. HFCS includes deposit-like instruments with non-deposit-taking corporations. These are classified as short term loans in FA.
F.3	**Debt securities**	HD1420	Bonds	Included	Conceptual differences are not known.
F.4	**Loans**	HD1710	Amount owed to household	Excluded	Not fully comparable, loans between households missing from FA in practice for most countries.
F.5	**Equity and investment fund shares**			Included	
F.511	Listed shares	HD1510	Publicly traded shares	Included	
F.512	Unlisted shares	HD1010	Investment in non-self-employment not publicly traded shares (ownership only as an investor or silent partner)	(Partly) Included dependent on adjusted concept	- In the HFCS, classification is based on the household's activity in the enterprise.

Table A1. Continued.

ESA 2010 code	FA/Instrument name	HFCS variable code(s)	HFCS variable	Adjusted concept	Remarks
F.519	Other equity	DA1140 (Sum of (HD080x) + HD0900	Investments in Self-Employment Businesses 1 – Sole proprietorship/ independent professional 2 – Partnership 3 – Limited liability companies 4 – Co-operative societies 5 – Nonprofit making bodies 6 – All other Forms (Spain) 7 – Unknown (not imputed)		- FA value includes assets that are classified as real wealth in the HFCS (value of self-employment businesses) and has to be reallocated to financial wealth. - The split between 'Unlisted shares' and 'Other equity' cannot be made in the survey. Investments in self-employed businesses could be included in 'Unlisted shares' or 'Other Equity'. - The value of sole proprietorships or partnerships are spread over the different instruments in FA if it is not considered as a separate legal entity (quasi-corporation). - In the HFCS, the value can be provided for the different legal status, although the legal status is not imputed in all countries. ("Unknown" category).
F.521	Money Market Fund shares/ units	HD1320c	Investments in mutual funds c – Funds predominantly investing in money market instruments	Included	Value dependent on fund type not imputed in every country. The breakdown by type of fund may not be available and only the total HD1330 is imputed in all countries. Hence the distinction between MMF and non-MMF funds may not be made in these countries.

Table A1. Continued.

ESA 2010 code	FA/Instrument name	HFCS variable code(s)	HFCS variable	Adjusted concept	Remarks
F.529	Non-MMF Fund shares/ units	HD1320x	a – Funds predominantly investing in equity b – Funds predominantly investing in bonds d – Funds predominantly investing in real estate e – Hedge funds f – Other fund types (specify)	Included	
F.6	**Insurance, pension and standardised guaranteed schemes**				
F.61	Nonlife insurance technical reserves	N/A	N/A	Excluded	Non-life included in N/A. Assets in nonlife (e.g., health insurance, term insurance) can be significant.
F.62	Life insurance and annuity entitlements	DA2109 (Sum of PF0920 over household members)	Voluntary pension/whole life insurance schemes	Included	
F.63	Pension entitlements	Sum of PF0700 over household members	Current value of all occupational pension plans that have an account	Excluded	It is not clear if defined benefit plans are included in this variable in the HFCS. Furthermore, pensions are prone to measurement problems in surveys.
F.64	Claims of pension funds on pension managers	N/A	N/A	Excluded	F.64–F.66 likely to be irrelevant for households.
F.65	Entitlements to nonpension benefits	N/A	N/A	Excluded	
F.66	Provision for calls under standardised guarantees	N/A	N/A	Excluded	

Table A1. Continued.

ESA 2010 code	FA/Instrument name	HFCS variable code(s)	HFCS variable	Adjusted concept	Remarks
F.7	**Financial derivatives**	HD1920	Other financial assets	Excluded	Financial derivatives are not a separate item in the HFCS and are included in 'Other financial assets'. Definition of 'Other accounts receivable/ payable' not comparable to 'Other financial assets', different definitions.
F.8	**Other accounts receivable/ payable**	HD1620	Managed accounts	Included	May be spread over the FA balance sheet of the household depending on set up of the management and dependent on the assets invested in. However, does not affect comparability of total financial assets.

7. References

Alvaredo, F., A. Atkinson, L. Chancel, T. Piketty, E. Saez, and G. Zucman. 2016. "Distributional National Accounts (DINA) guidelines: Concepts and methods used in WID.world." *WID Working Paper Series*, No. 2016/2. Available at: https://wid.world/document/dinaguidelines-v1/ (accessed February 2019).

Alvaredo, F., L. Chancel, T. Piketty, E. Saez, and G. Zucman. 2017. "Global inequality dynamics: New findings from WID.world." *NBER Working Paper*, No. 23119. Doi: https://doi.org/10.3386/w23119.

Andreasch, M., P. Fessler, and P. Lindner. 2013. "Linking Microdata and Macrodata on Austrian Household Financial Wealth Using HFCS and Financial Accounts Data." *Statistiken*, Special Issue, Oesterreichische Nationalbank, 14–23. Available at: https://www.oenb.at/dam/jcr:048c539b-0f03-480e-992c-db2948c68416/stat_2009_q1_gesamt_tcm14-96306.pdf (accessed February 2019).

Antoniewicz, R. 2000. "A Comparison of the Household Sector from the Flow of Funds Accounts and the Survey of Consumer Finances." *Federal Reserve Board of Governors Survey of Consumer Finances Working Paper*. Available at: https://www.federalreserve.gov/econresdata/scf/files/antoniewicz_paper.pdf (accessed February 2019).

Bach, S., A. Thiemann, and A. Zucco. 2015. "The Top Tail of the Wealth Distribution in Germany, France, Spain, and Greece." Discussion Paper 1502, DIW Berlin. Available at: https://www.diw.de/sixcms/detail.php?id=diw_01.c.513263.de (accessed February 2019).

Bover, O., E. Coronado, and P. Velilla. 2014. "The Spanish survey of household finances (EFF): description and methods of the 2011 wave." *Banco de España occasional Paper* 1407. Available at: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/14/Fich/do1407.pdf (accessed February 2019).

Bricker, J., A. Henriques, J. Krimmel, and J. Sabelhaus. 2016a. "Measuring Income and Wealth at the top using administrative and survey data." *Brookings Papers on Economic Activity,* Spring: 261–312. Available at: https://www.brookings.edu/wp-content/uploads/2016/03/brickertextspring16bpea.pdf (accessed February 2019).

Bricker, J., A. Henriques, J. Krimmel, and J. Sabelhaus. 2016b. "Online Appendix to Measuring Income and Wealth at the top using administrative and survey data." Available at: https://www.brookings.edu/wp-content/uploads/2016/03/brickerappendixspring16bpea.pdf (accessed February 2019).

Chakraborty, R. and S. Waltl. 2018. "Missing the wealthy in the HFCS: Micro problems with macro implications." ECB Working Paper Series, 2163. Available at: https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2163.en.pdf? (accessed February 2019).

Davies, J.B. and A. Shorrocks. 1999. "The Distribution of Wealth." In *Handbook of Income Distribution: Volume 1*, edited by A.B. Atkinson and F. Bourguignon, 605–675. North-Holland, Amsterdam.

Dettling, L.J., S.J. Devlin-Foltz, J. Krimmel, S.J. Pack, and J.P. Thompson. 2015. "Comparing Micro and Macro Sources for Household Accounts in the United States: Evidence from the Survey of Consumer Finances." *Finance and Economics Discussion Series* 2015-086. Washington: Board of Governors of the Federal Reserve System. Doi: http://dx.doi.org/10.17016/FEDS.2015.086.

Dolan, K.A. 2016. "Methodology: How We Crunch the Numbers." Forbes www-sides. Available at: https://www.forbes.com/sites/kerryadolan/2012/03/07/methodology-how-we-crunch-the-numbers/ (accessed February 2019).

ECB. 2012. "HFCS Core Variables Catalogue Wave I." Available at: https://www.ecb.europa.eu/home/pdf/research/hfcn/core_output_variables.pdf (accessed February 2019).

ECB. 2013. "The Eurosystem Household Finance and Consumption Survey: Methodological Report for the First Wave." *European Central Bank Statistics Paper Series*, 1. Available at: https://www.ecb.europa.eu/pub/pdf/other/ecbsp1en.pdf (accessed February 2019).

Eckerstorfer, P., J. Halak, J. Kapeller, B. Schütz, F. Springholz, and R. Wildauer. 2016. "Correcting for the Missing Rich: An Application to Wealth Survey Data." *Review of Income and Wealth* 62(4): 605–627. Doi: https://doi.org/10.1111/roiw.12188.

European System of Accounts 1995 (ESA95). Council (EC) Regulation No 2223/96 of 25 June 1996 on the European system of national and regional accounts in the Community, Official Journal of the European Union L 310/1-469. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1996R2223:20030807:EN:PDF (accessed February 2019).

European system of accounts (ESA2010). Eurostat/European Commission, Publications Office of the European Union, Luxembourg 2013. Available at: https://ec.europa.eu/eurostat/documents/3859598/5925693/KS-02-13-269-EN.PDF/44cd9d01-bc64-40e5-bd40-d17df0c69334 (accessed February 2019).

Gabaix, X. and R. Ibragimov. 2011. "Rank -1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents." *Journal of Business and Economic Statistics* 29: 24–39. Doi: https://doi.org/10.1198/jbes.2009.06157.

Henriques, A.M. and J.W. Hsu. 2014. "Analysis of Wealth Using Micro and Macro Data: A Comparison of the Survey of Consumer Finances and Flows of Funds Accounts." In *Measuring Economic Sustainability and Progress*, edited by D.W. Jorgenson, J.S. Landefeld, and P. Schreyer. University of Chicago Press.

IMF/FSB report to the G-20 Finance Ministers and Central Bank Governors. Available at: http://www.financialstabilityboard.org/publications/r_091107e.pdf (accessed February 2019).

Kavonius, I.K. and J. Honkkila. 2013. "Reconciling Micro and Macro Data on Household Wealth: A Test Based on Three Euro Area Countries." *Journal of Economic and Social Policy* 15(2). Article 3. Available at: https://search.informit.com.au/documentSummary;dn=592344712313054;res=IELBUS (accessed February 2019).

Kavonius, I.K. and V.M. Törmälehto. 2010. "Integrating Micro and Macro Accounts – The Linkages between Euro Area Household Wealth Survey and Aggregate Balance Sheets for Households." IARIW 31st General Conference, St Gallen, Switzerland, August 22–28, 2010. Available at: http://www.iariw.org/papers/2010/7akavonius.pdf (accessed February 2019).

Kennickell, A.B. 2008. "The Role of Over-sampling of the Wealthy in the Survey of Consumer Finances." *Irving Fisher Committee Bulletin* 28: 403–408. Available at: https://www.bis.org/ifc/publ/ifcb28.pdf#page=409 (accessed February 2019).

Kennickell, A.B. and R.L. Woodburn. 1999. "Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth." *Review of Income and Wealth* 45(2): 193–215. Doi: https://doi.org/10.1111/j.1475-4991.1999.tb00328.x.

Pérez-Duarte, S., C. Sánchez-Muñoz, and V.M. Törmälehto. 2010. "Re-weighting to reduce unit non-response bias in household wealth surveys: a cross-country comparative perspective illustrated by a case study." In *Conference on European Quality in Statistics*, Helsinki. Available at: https://www.ecb.europa.eu/home/pdf/research/hfcn/WealthSurveys.pdf (accessed February 2019).

Piketty, T. 2014. *Capital in the Twenty-First Century*. Cambridge and London: Harvard University Press.

Piketty, T., E. Saez, and G. Zucman. 2018. "Distributional national accounts: methods and estimates for the United States." *The Quarterly Journal of Economics* 133(2): 553–609. Doi: https://doi.org/10.1093/qje/qjx043.

Rubin, D.B. 2004. "Multiple imputation for nonresponse in surveys." John Wiley & Sons.

Saez, E. and G. Zucman. 2016. "Wealth inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *The Quarterly Journal of Economics* 131(2): 519–578. Doi: https://doi.org/10.1093/qje/qjw004.

Stiglitz, J.E., A. Senand, and J.P. Fitoussi. 2009. "*Report by the Commission on the Measurement of Economic Performance and Social Progress*." Available at: https://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (accessed February 2019).

System of National Accounts 1993 (SNA93). Commission of the European Communities, International Monetary Fund, United Nations, World Bank, Brussels/Luxembourg,

New York, Paris, Washington D.C., 1993. Available at: https://unstats.un.org/unsd/nationalaccount/docs/1993sna.pdf (accessed February 2019).

System of National Accounts 2008 (SNA2008). Commission of the European Communities, International Monetary Fund, United Nations, World Bank, New York 2009. Available at: https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf (accessed February 2019).

Tissot, B. 2015. "Closing Information Gaps at the Global Level – What Micro Data Can Bring." Irving Fisher Committee Workshop, Warsaw 14–15 December 2015. Available at: https://www.bis.org/ifc/events/ws_micro_macro/tissot_paper.pdf (accessed February 2019).

Vermeulen, P. 2016. "Estimating the Top Tail of the Wealth Distribution." *American Economic Review* 106(5): 646–650. Doi: http://dx.doi.org/10.1257/aer.p20161021.

Vermeulen, P. 2018. "How Fat is the Top Tail of the Wealth Distribution?" *Review of Income and Wealth* 64(2): 357–387. Doi: https://doi.org/10.1111/roiw.12279.

Received February 2018
Revised May 2018
Accepted August 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 67–92, http://dx.doi.org/10.2478/JOS-2019-0004

Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey

Leandro D'Aurizio[1] *and Giuseppina Papadia*[2]

We examine two sources of bias for the Bank of Italy's panel business survey of Industrial and Services Firms:

1) the bias caused by panel attrition; and

2) the bias created by delays in the distributional data on the reference population, needed for computing the survey weights.

As for the first source of bias, the estimates strongly dependent on big firms' values are less affected by panel attrition than those representing firms' average behavior, independent of their sizes. Positive economic results make it easier to enroll new firms in the survey, in order to replace firms dropping out because of bad economic performances. However, the economic results of new entrances become more aligned to those of the population, once they enter the sample.

A very different result emerges for the second source of bias, since, when the population size is highly variable, the information delays produce a bias for the estimates influenced by the contribution of great firms, but the effect is negligible for the estimates not dependent on firm size.

Key words: Business surveys; panel samples; panel attrition; administrative data; auxiliary information.

1. Introduction

Business surveys are often conducted by using a panel sample, with estimates that can be affected by panel attrition. Since these estimates are representative of a reference population by means of survey weights, they can be biased because the weights do not fully take into account the evolution of the reference population. Our article aims to evaluate these two sources of bias by relying on auxiliary information from administrative sources. We will understand the effects of panel attrition by using an integrated archive, matching survey data with financial-statements indicators, available also for the years when the panel units are absent from the sample. We will assess the effects of the second source of bias by fully exploiting the population information.

Measuring ratios is a typical utilization of business surveys. They can be either ratios between a variable at time t and the same variable at time $t - 1$, in order to measure its relative change over time, or they can be ratios between two different variables. For them,

[1] Italian Authority for the Supervision of the Insurance Sector (IVASS), Research and Data Management Directorate, via del Quirinale 21, 00187 Rome, Italy. Email: leandro.daurizio@ivass.it.

[2] Bank of Italy, Economics and Statistics Department, via Nazionale 91, 00181 Rome, Italy. Email: giuseppina.papadia@bancaditalia.it.

Acknowledgments: We thank Andrea Brandolini, Luigi Cannari, Pierluigi Conti, Giovanni D'Alessio, Stefano Iezzi, Giuseppe Ilardi, Alfonso Rosolia and Giordano Zevi for their insightful suggestions. The views expressed in this article are those of the authors and do not imply any responsibility of their institutions.

two classes of estimators can be computed: the first one comprises what we define as simple estimates, representative of firms' average behavior, not influenced by the largest units in the sample. The second class includes what we define as aggregate estimates, which tend to be heavily influenced by the big enterprises belonging to the sample, especially if they are over-represented, as in the case of Neyman's samples.

We focus on the Bank of Italy's Survey of Industrial and Service-sector firms (*Invind*, from now on) conducted every year with a panel of around 4,000 enterprises (Bank of Italy 2014), representative of the population of industrial and nonfinancial service-sector firms with at least 20 employees.

The first source of bias we examine arises from firms participating in a survey edition and dropping out from the following one, without that being planned in advance. The term normally used for this phenomenon is panel attrition (Martin et al. 2001). We analyze the attrition effects by using an archive of financial-statement data for the whole reference population. It provides indicators for all the panel units for every survey year, even if they are missing from the sample in several survey editions. We can therefore measure whether, for each survey edition, new entrants and dropouts are different and the main determinants of the propensity to enter and leave the sample.

The second source of bias we assess is caused by the delays of the distributional data on the reference population, required to compute the survey weights. The effect of this lag can be measured for the least recent survey editions, for which complete information on the population is available. We can therefore assess how much the bias of the usual estimates derives from out-of-date population information.

We briefly anticipate the main results of our article. The effects of panel attrition on the aggregate estimates are small, since their values depend on the data of the largest companies, which tend to participate in the survey more regularly than smaller firms. The official estimates, regularly analyzed in the Bank of Italy's reports are of this kind. On the other hand, the smaller firms tend to participate in the survey more erratically and accordingly this fact makes it necessary to carefully interpret the simple estimates. On the contrary, the delays in the updating of the reference population are a source of bias for the aggregate estimates, when they are weighted to be representative of the whole reference population. More precisely, they tend to be biased when the population size is highly unstable, with the bias virtually disappearing for the simple estimates.

The article is organized as follows. Section 2 introduces panel attrition. Section 3 describes the specific features of the panel attrition found in *Invind*. Section 4 describes the indexes we use to assess the effect of panel attrition. Section 5 explains the data integration process. Section 6 evaluates the effects of panel attrition effects on *Invind* and Section 7 analyzes the consequences of the delayed updating of the distribution of the reference population. Section 8 sums up the main results and proposes some solutions to manage the problems highlighted in the article.

2. Panel Attrition: An Overview

Panel surveys use the same sample units for repeated survey occasions. This choice enables researchers to understand transition patterns (Fabbris 1989). An obvious operational advantage of panels over repeated independent samples is that the sample is

selected once before the first survey edition. In most cases, the original panel undergoes an attrition process after the first survey occasion and its composition accordingly changes over time. The statistical literature has widely explored the cases of panel attrition created by some sample units either refusing to participate in the following occasions or exiting the reference population. The units that leave the panel after a given survey occasion are routinely replaced by other ones, that either have been in the panel in previous waves and reenter in this occasion, as replacements of other units, or they are enrolled in the panel for the first time (some of them may even be new entries into the reference population).

If the attrition process decreases the sample size, the standard error of the survey estimates automatically increases. The estimates remain unbiased if the attrition is completely random (Little and Rubin 2002). On the contrary, if the attrition depends on some of the variables of interest for the survey, it becomes a source of bias.

There are some solutions to attenuate these drawbacks. Typically, the units leaving the sample are replaced by others, with observable characteristics (the same used in the survey design) quite similar to those of the replaced units. A more complex solution is to set up panel rotations that periodically discard a part of the panel and make place for new units, that can also realign the sample to the changes in the reference population occurred since the creation of the panel (Trivellato 1999). Rotating panels also spread the response burden and therefore indirectly reduce the attrition caused by it (Ardilly and Lavallée 2007). The method is widely applied in surveys of individuals and households.

These measures maintain the initial sample size and keep the precision of the estimates close to the planned levels. If the hypothesis of attrition totally at random is violated, the replacements of the units leaving the panel cannot totally eliminate the bias, since the units leaving and entering the panel may differ according to characteristics not included in the survey design. The bias arises when such characteristics are correlated with the variables of interest.

Among the recent contributions to the statistical literature on panel attrition, the utilization of many techniques for imputing data, missing because of attrition, can provide useful clues regarding the direction of the bias. For instance, Black et al. (2007) use an array of imputation methods, ranging from simple mean imputation to more complex Bayesian resampling techniques, to reconstruct the missing values for a sample of UK data on car traffic. They measure the variability of the results obtained, in order to assess the bias from missing data and its influence on the estimates.

Deng et al. (2013) use the waves of the Survey on Income and Program Participation, regularly conducted on a representative panel sample of US households, and propose to use a series of refreshment samples, composed by new, randomly selected, respondents, as an auxiliary external source. The new respondents answer the same questionnaire used for the main sample. The differences between the answers of the two respondent sets are used to correct for the bias.

Auxiliary information can also be used as a correction factor of the original weighting system, in order to offset the effects of panel attrition. For example, Afonso (2015) corrects the weights to compensate for the missing data created by attrition in a panel of bank microdata that follows the latest trends of the net interest margin of the banking systems of the 15 major countries of the euro area. The revised weights are used in the

context of Generalized Methods of Moments estimation procedures to produce consistent estimates better aligned to the predictions of economic theory.

3. Panel Attrition in the Invind Survey

We concentrate on the attrition observed between two consecutive waves of the *Invind* survey relative to the years $t-1$ and t, that we indicate with I_{t-1} and I_t (we do not study the effects of attrition on the longitudinal estimates based on the panel). Some firms participate in I_{t-1}, but not in I_t (we refer to them as "dropouts") and are accordingly replaced by other firms that did not participate in I_{t-1} (we call them "new entrants"). We use the term "stayers" to indicate the firms participating in both I_{t-1}, and I_t. We use these three terms looking at just two consecutive panel waves, without taking into account what happens in the other ones. We differ here from the standard utilization of these three terms in the literature, indicating the units that respectively drop out of the panel definitively, enter it for the first time, or regularly participate in all the survey editions.

In the single cross-sections, the sample is representative of the cross-sectional population with the help of replacement rules that substitute every dropout with a new entrant, having its head office in the same Italian region, together with economic activity and number of employees as close as possible to those of the dropout.

We consider all the waves available from 2002 until 2013. The wave relative to 2002 was the first with the current reference population (composed by the firms with at least 20 employees belonging to the sectors of nonconstruction industry and nonfinancial private services) and the current sample size of 4,000–4,200 firms (Bank of Italy 2005).

The interviews relative to the wave for year t take place in the first four to five months of the following year $t+1$. The survey collects the values for the main variables of interest (employment, turnover and investment) for the years $t-1$, t and $t+1$ (this last value is a forecast for the current year). The changes for the year t relative to $t-1$ and for the year $t+1$ relative to t are accordingly computed by using only one survey wave. If extraordinary events (such as mergers, acquisitions or splits) modify the structure of a panel firm between $t-1$ and $t+1$, its data cannot be directly used to compute the average variations and require a special treatment. The data are included in the estimates only if they refer to a set of plants and workers fully comparable over the three years $t-1$, t and $t+1$. The comparability is obtained either by anticipating the extraordinary events at the beginning of $t-1$, or by postponing them at the end of $t+1$.

Invind weights are cross-sectional. For every survey edition, they make the sample representative of the reference population within strata formed by the combinations of six class sizes (in terms of average number of employees: 20–49, 50–99, 100–199, 200–499, 500–999, 1000–4999, 5000 and over) and eleven sectors of economic activity. A successive post-stratification makes the sample representative of the population also at the geographical level: there are 48 post-strata: north-west, north-east, center, south and islands, referred to the firm headquarters' location, combined with size classes and aggregate economic sectors (Table S1 in Supplementary material). A unique set of survey weights is used for each survey edition, relative to the year t.

Many firms are dropouts or new entrants during the years 2002–2013 (Table 1). On average, 20% of the units in a wave drop out from the following one and are replaced

Table 1. Invind survey: yearly attrition rates (2002–2013).

Year	Total number of firms in the sample	Stayers		Dropouts							New entrants			
		Firms participating in the year's wave and still present in the following wave		Firms participating in the year's wave and exiting the following wave							Firms participating in the year's wave and not present in the previous wave		*of which: firms participating in the past waves*	
						Distribution of the firms exiting the following wave (%)								
		Per cent	Number	Per cent	Number	Firms still active with 20 employees and over	Firms still active with less than 20 employees	Firms involved in mergers, acquisitions, contributions, transfers and splits	Firms no longer active	Total	Per cent	Number	*Per cent*	*Number*
2002	3,969	78.2	3,105	21.8	864	69.9	7.1	2.9	20.1	100.0	–	–	–	–
2003	4,135	78.4	3,240	21.6	895	74.2	9.3	2.0	14.5	100.0	24.9	1,030	*2.6*	*109*
2004	4,226	80.0	3,382	20.0	844	74.1	7.2	2.1	16.6	100.0	23.3	986	*3.4*	*142*
2005	4,386	79.8	3,498	20.2	888	77.5	8.2	2.0	12.3	100.0	22.9	1,004	*4.4*	*195*
2006	4,252	80.3	3,415	19.7	837	75.5	9.0	1.6	14.0	100.0	17.7	754	*4.7*	*199*
2007	4,063	79.2	3,217	20.8	846	74.6	10.3	1.7	13.5	100.0	15.9	648	*5.3*	*214*
2008	3,952	80.3	3,175	19.7	777	78.2	8.1	2.3	11.3	100.0	18.6	735	*6.8*	*270*
2009	3,921	80.6	3,162	19.4	759	78.0	7.9	2.0	12.1	100.0	19.0	746	*7.4*	*291*
2010	3,937	82.5	3,248	17.5	689	78.4	8.9	0.4	12.3	100.0	19.7	775	*8.0*	*313*
2011	4,120	82.4	3,396	17.6	724	77.6	7.3	1.7	13.4	100.0	21.2	872	*8.8*	*361*
2012	4,213	82.1	3,460	17.9	753	74.1	8.1	2.0	15.8	100.0	19.4	817	*8.6*	*362*
2013	4,215	–	–	–	–	–	–	–	–	–	17.9	755	*11.8*	*498*
Total		**80.3**		**19.7**		**75.6**	**8.3**	**1.9**	**14.2**	**100.0**	**20.0**		***6.5***	
of which:														
Non-construction industry		81.1		18.9							19.0			
Non-financial private services		78.3		21.7							23.0			
Average firm size *(number of employees)*		**390**		**222**							**303**			

Source: Bank of Italy's Invind business survey.

by a slightly higher number of units (the sample size increased by 6.5% in the time span examined). An average 75% of the dropouts still belong to the reference population of firms with at least 20 employees, 14% are no longer operational and 10% become not eligible for further participation, either because their workforce drops below the 20-employee threshold or, more rarely, because of events such as mergers, acquisitions, and so on. The stayers' average size is 390 employees, higher than that of the new entrants (303), that is in turn higher than dropouts' average size (222). The last two columns of Table 1 show that, on average, a third of the new entrants participated in past waves before I_{t-1}. The shares of dropouts and new entrants are quite stable throughout the period examined. The economic crisis that began in 2009 and a revision of the survey operations in the years 2006–2008 did not significantly alter this pattern.

Looking at the firms' economic sectors, the attrition is stronger for firms of the nonfinancial private services (Table 1, last row) that became a part of the reference population only since 2002. The interplay of three factors explains this result:

1. it takes time to create a stable panel participation, because the questionnaire is difficult to complete,
2. service-sector firms tend to outsource most of the budgeting and accounting tasks needed to complete essential parts of the questionnaire and this discourages their regular participation in the survey,
3. firms' transformations (by mergers, acquisitions, contributions, transfers and splits) naturally decrease the propensity to participate in business surveys and tend to affect service-sector firms more frequently (Bank of Italy 2015).

4. Simple and Aggregate Indexes

An array of variables can be measured over a firm, either dimensional (such as profits, turnover, investment and number of employees) or dimensionless (e.g., ratios like Return on Assets and Return on Equity, indicated with the acronyms ROA and ROE from now on, extensively used to evaluate firms' performances) and both can be summarized through indexes.

For both kinds of variables, we can follow how their averages vary over the years. Let us indicate with i a unit of the sample of size n_t. The indicators can take two different forms, which we define as simple and aggregate indexes. For both forms, they can be either weighted with the cross-sectional survey weights or can be left unweighted.

The units contribute equally to the simple indexes (Table 2, lower part), regardless of their different sizes, while their contribution to the aggregate indexes (Table 2, upper part) takes place according to their relative sizes. It can be easily shown that the relative size is the value of the variable at $t - 1$ divided by its total (we report below the expression for the unweighted aggregate index):

$$\frac{\sum_{i=1}^{n_t} y_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1}} = \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \frac{y_{i,t-1}}{\sum_{i=1}^{n_t} y_{i,t-1}} \tag{1a}$$

Table 2. Key performance indicators of firms.

	Aggregate indexes	
	Dimensional variables[(a)]	Dimensionless ratios[(b)]
Unweighted	$\frac{\sum_{i=1}^{n_t} y_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1}}$	$\frac{1}{\sum_{i=1}^{n_t} z_{i,t}} \sum_{i=1}^{n_t} x_{i,t} z_{i,t}$
Weighted	$\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}}$	$\frac{1}{\sum_{i=1}^{n_t} z_{i,t} w_{i,t}} \sum_{i=1}^{n_t} x_{i,t} z_{i,t} w_{i,t}$
	Simple indexes	
	Dimensional variables[(a)]	Dimensionless ratios[(b)]
Unweighted	$\frac{1}{n_t} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}}$	$\frac{1}{n_t} \sum_{i=1}^{n_t} x_{i,t}$
Weighted	$\frac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} w_{i,t}$	$\frac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} x_{i,t} w_{i,t}$

(a) E.g.: turnover and profit. – (b) E.g.: ROA and ROE.

Equation (1a) can be easily generalized to the weighted aggregate index in the following way:

$$\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} = \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} \tag{1b}$$

For a dimensionless ratio $x_{i,t}$ the scale factor is a variable $z_{i,t}$ positively correlated with firm size (generally number of employees or turnover), divided by its total. We will use the turnover for our computations.

5. Using External Sources to Measure Invind Panel Attrition

5.1. The Main Issues

The high attrition levels shown in Section 3 require to assess whether, for every cross-section, the economic performances of the firms entering the cross-section, but not present in the previous one (new entrants) and those of the firms absent in the following cross-section (dropouts) are different. We measure the economic performances in terms of turnover changes, profits changes, ROA and ROE, using the indexes defined in Section 4.

It is also relevant to study the propensity to enter or leave a cross-section, with the first propensity requiring firm-level data for new entrants also for the year prior to their entrance, not directly available from the survey.

We therefore face two problems requiring a data integration process:

1) only turnover is collected in the survey, but not profits, ROA and ROE;
2) data should also be available for the years when the units are absent from the sample.

5.2. The Data Integration Process

5.2.1. The New Archive

Since all the panel firms are limited companies, apart from a negligible number of partnerships (less than 0.1% on average), we integrate the *Invind* data with the *Cerved* archive, a data warehouse for all the Italian limited companies' financial statements filed since 1993. We use the firm VAT number, available on both sources, to exactly match the survey data with the corresponding financial figures. The matching fails whenever the VAT number is missing in *Invind* or it does not find, when it is present, a corresponding VAT in *Cerved* because of errors in one or both sources.

5.2.2. The Matching Quality

A first clue of the quality of the matching is the high percentage of matched units (on average higher than 90%, Table 3), that remains stable (as shown in Figure 1) within the categories of the variables used in the stratification and post-stratification steps. This latter result implies that the survey design also keeps under control the bias caused by using a sample smaller than the original one.

For the average number of employees and the turnover levels, found in both sources, we also examine the individual differences between the two corresponding values from the two sources. If we indicate with $inv_{j,i,t}$ and $cer_{j,i,t}$ the values derived respectively from *Invind* and *Cerved,* relative to the *j*-th variable for the *i*-th matched unit in the year *t*, the size of the absolute difference $\left|inv_{j,i,t} - cer_{j,i,t}\right|$ would depend too much on the unit of measure. We eliminate this effect by using a standardized absolute difference $sad_{j,i,t}$, expressed as:

$$sad_{j,i,t} = \frac{\left|inv_{j,i,t} - cer_{j,i,t}\right|}{\frac{inv_{j,i,t} + cer_{j,i,t}}{2}} 100 \tag{2}$$

Table 3. Percentage of firms in Invind annual surveys matched with financial-statement archives (2002–2013).

	Nonconstruction industry firms	Nonfinancial private services firms	Yearly total
2002	87.8	89.6	88.2
2003	87.6	89.1	88.0
2004	88.6	91.8	89.4
2005	92.4	93.0	92.5
2006	93.3	92.6	93.1
2007	93.1	93.4	93.2
2008	92.1	94.2	92.7
2009	93.2	93.5	93.2
2010	93.4	94.8	93.8
2011	93.8	95.1	94.2
2012	94.3	94.6	94.4
2013	90.3	88.9	89.9
Total	**91.6**	**92.6**	**91.9**

Source: Bank of Italy's Invind business survey and Cerved archive.

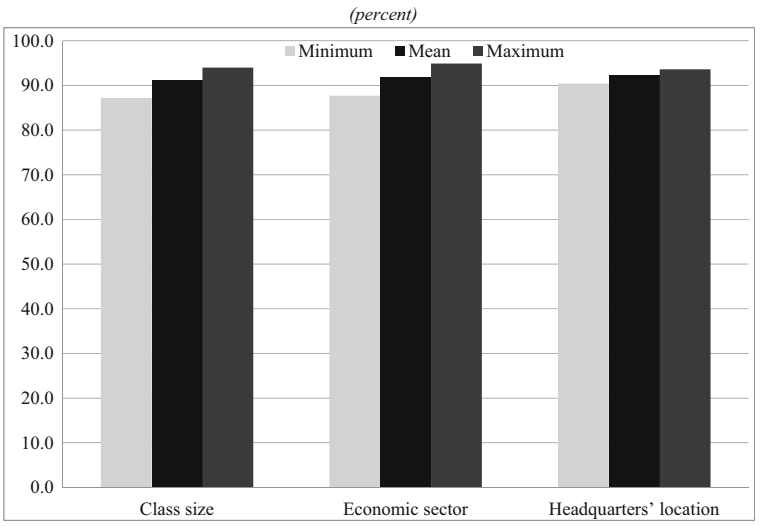

Fig. 1. Minimum, mean and maximum of the average percentage of matched Invind units for all the categories of class size, economic activity and headquarters' location. Source: Bank of Italy's Invind business survey and Cerved archive.

We discard from the analysis the firms affected by structural changes (5.4% for the years 2002–2013), for which the indicator is naturally big, because in such cases the *Invind* figures are adjusted and not comparable with the corresponding ones from *Cerved* (see Section 3 for details on the adjustment method).

The percentage of observations with same values in the two sources are 23.8% (Table 4) for the average number of employees and 11.0% respectively for the turnover levels. The size of the differences is limited for turnover levels, since the median value of the *sad* is 1.07% for all the matched observations (including those for which the turnover values from *Invind* and *Cerved* are identical) and the indicator is negatively correlated with the number of employees. For the average number of employees the indicator is higher, with its median value equal to 3.08% for all the matched observations, but it still decreases with firm size. Since we will use only a categorization of this variable in our developments, the risk generated from having different employment values in *Invind* and *Cerved* for the same unit is contained.

A remaining concern is that the signs of the differences might follow systematic patterns. Leaving aside the cases where $inv_{j,i,t} = cer_{j,i,t}$, we probe into the issue by looking at the successions of cases for which $inv_{j,i,t} > cer_{j,i,t}$, or $inv_{j,i,t} < cer_{j,i,t}$, in order to verify whether they are random or follow systematic patterns. First we create a binary variable:

$$d_{j,i,t=}\begin{cases} 0, & if\ inv_{j,i,t} < cer_{j,i,t} \\ 1, & if\ inv_{j,i,t} > cer_{j,i,t} \end{cases}$$

We then randomly sort the observations and finally we measure the nonparametric Wald-Wolfowitz test (Hollander and Wolfe 1973) separately for the average number of employees and the turnover each year from 2002 to 2013. The high sample size allows us to use the asymptotic normal distribution to compute the test. With a significance level of 1%, we cannot reject the null hypothesis of randomness of the successions of 0 and 1 for

Table 4. Standardized absolute difference (sad) between Invind and Cerved figures for the matched observations (2002–2013).

Average number of employees				
	Percentiles			
	25%	50%	75%	90%
Class size				
20–49	2.30	7.69	210.26	216.00
50–99	1.05	2.99	9.27	88.66
100–199	0.00	1.72	5.94	20.62
200–499	0.00	0.98	3.83	13.17
500–999	0.00	0.70	3.33	11.73
1,000 and over	0.00	0.57	3.29	14.80
Total	**0.22**	**3.08**	**13.08**	**210.00**
Number of matched observations	42,940			
% of matched observations with $sad = 0$	23.8			
Turnover				
	Percentiles			
	25%	50%	75%	90%
Class size				
20–49	0.18	1.24	4.78	14.20
50–99	0.18	1.12	4.20	12.70
100–199	0.10	0.95	3.90	11.45
200–499	0.06	0.81	3.56	11.55
500–999	0.03	0.76	3.89	12.90
1,000 and over	0.02	1.06	5.32	15.76
Total	**0.13**	**1.07**	**4.27**	**13.07**
Number of matched observations	42,846			
% of matched observations with $sad = 0$	11.0			

Source: Bank of Italy's Invind business survey and Cerved archive.

$d_{j,i,t}$ respectively in nine and eleven of all the 12 years considered for the turnover and for the average number of employees (results of the test are available on request).

We therefore conclude that the quality of the matching is acceptable.

5.2.3. The Imputation of the Employment Levels

A necessary step is the imputation of the average number of employees for the firms of the *Cerved* archive without this figure. We need the workforce levels for all the firms belonging to the *Invind* reference population (formed by the firms with at least 20 employees operating in the economic sectors of interest) in order to model the probability to enter the cross-sectional sample, since a categorization of this figure is a control variable in the model.

We use the predicted values of a linear model estimated by Ordinary Least Squares (OLS) to impute the missing data. We estimate the model only for the firms with turnover greater than a threshold represented by the lowest turnover found in the survey minus 20%.

This restriction has two aims: 1) preventing the model from being excessively influenced by smaller firms; 2) using units with similar probabilities of having missing number of employees, since firms below the threshold tend to be more affected by the problem. The dependent variable in the model is the average workforce level for the year, the covariates include some economic indicators (ROA, ROE, turnover), together with economic sector of activity and geographical location of firm's head office.

We select the best model within a set of possible specifications according to the following criteria:

- the covariates should not have a high frequency of missing values;
- the specification should be reasonable according to basic economic theory, even though no causal modeling is attempted;
- the fit to the data should be good (in terms of *R-square*);
- estimated totals of firms and employees for the most relevant classification cells should be close to those of the official aggregate evidence available from the Statistical Archive of Active Enterprises (ASIA), provided by the Italian National Statistical Institute (Istat 2014).

We estimate the selected model separately within cells (Table S2 in Supplementary material) formed by the combination of 20 analytical economic activities (instead of eleven ones used in the stratification) and geographical locations of the firm's headquarters (with the same level of detail as in the post-stratification). Using separate models with the same specification substantially improves the values of the adjusted R-square (it ranges between 0.7 and 0.9 for the various cells), with respect to the alternative of using the economic activity and the geographical location as covariates in a unique regression, with common parameters for the quantitative regressors. We report the model equation below:

$$Employees_{t,j,i} = \beta_{0t,j} + \beta_{1t,j} turn_{t,j,i} + \beta_{2t,j} turn^2_{t,j,i} + \beta_{3t,j} roa_{t,j,i} + \beta_{4t,j} roe_{t,j,i} + \varepsilon_{t,j,i} \quad (3)$$

where *t*, *j*, *i* and *turn* respectively indicate year, cell, individual firm and turnover. According to economic theory, labor cost is a better predictor than turnover for the number of employees in specifications like those shown in Equation (3). Unfortunately, we cannot use them because they are missing in 29% of cases. Firms' turnover is however a reasonable proxy, because it is highly correlated with labor costs within the cells where we estimate the model (the correlation coefficient is 0.85 on average). By using this estimate to impute the missing values for the number of employees, we obtain distributions of enterprises and employees (for the firms with at least 20 employees) similar to those of the official sources for this population (Table S3 in Supplementary material).

We use multiple imputation to estimate the variability of the imputation process by creating ten independent replications of the model predictions, each obtained by adding to the model prediction a random drawing from the residual distribution.

6. Enterprises' Performances and Panel Attrition in the Invind Survey

6.1. *First Evidence of the Attrition Effects*

For every survey edition, we calculate the simple and aggregate indexes presented in Table 2, weighted with the survey weights, for turnover and profit changes, ROA and

ROE, over three groups: new entrants (absent in the previous edition), future dropouts (absent in the next edition) and the rest of the sample. We express them as percent relative changes by subtracting one and multiplying the result by one hundred.

The effect of panel attrition is clearly visible for the simple indexes that are greater for new entrants than for future dropouts (Figure 2). This is true on average over the whole period, as well as for all the single years, with the exception of 2006, 2010, and 2012 for profit changes. For the aggregate index, the distinction between new entrants and future dropouts is less clear-cut, since the indexes relative to the latter ones are higher than those for the former ones for four years of the turnover changes series, five years of the ROA series and again for four years of the series of profit changes. Only for ROE is the index for new entrants always greater than that for future dropouts, even if, on average, the two are slightly closer than in the case of the simple index (the average distance is 4.8 points against 5.0).

It is now relevant to determine whether the differences discussed above are statistically significant. With this aim, for every survey edition, we separately regress each of the four indicators considered on a binary dummy identifying the two groups of interest, controlling at the same time for the survey design variables.

For every year t, we estimate the following linear model:

$$y_{t,i} = \beta_{0t} + \beta_{1t} d_{t.i} + \underline{\beta}_{\underline{t}} \underline{X}_{\underline{t,i}} + \varepsilon_{t,i} \tag{4}$$

where $y_{t,i}$ is one of the four indicators considered, $d_{t,i}$ is the binary dummy indicating whether the unit is a new entrant or a future dropout and $\underline{X}_{t,i}$ is an array of firm-level characteristics (firm size, economic sector of activity, headquarters' geographical location). We also estimate a synthetic version of the model (4) on the pooled data set and we finally replicate the regressions separately on the firms with 20–99 employees and on those with 100 employees and over.

In the year-by year regressions over the firms with 20 employees and over, the coefficients relative to the dummy are almost always positive and significant for turnover changes, ROE and ROA (Table 5). For profit changes, the effect of the dummy is always positive, but is statistically significant only for one year. The dummy coefficient is always positive and highly significant also in the corresponding regressions over the pooled data. The regressions on the two separate groups of firms with 20–99 employees and 100 employees and over reveal that, for the group of bigger firms, the differences between new entrants and future dropouts attenuate and the shares of new entrants and future dropouts decrease.

These results are consistent with the descriptive evidence of Figure 2, since they show that the economic performances of the firms not regularly participating in the survey tend to be more similar when firm size increases, even after conditioning with other observable characteristics.

6.2. *Modeling Panel Attrition*

Modeling the propensities to enter a given survey wave and to drop out of it helps us explain what drives this behavior. For this aim, we estimate a *logit* model over a pooling of all the observations. We use the waves for the years from 2003 to 2013 to analyze the new entrants, those from 2002 to 2012 for the dropouts. ROE, ROA, turnover and profit changes are our main covariates, to which we add dummies for the years, to take into

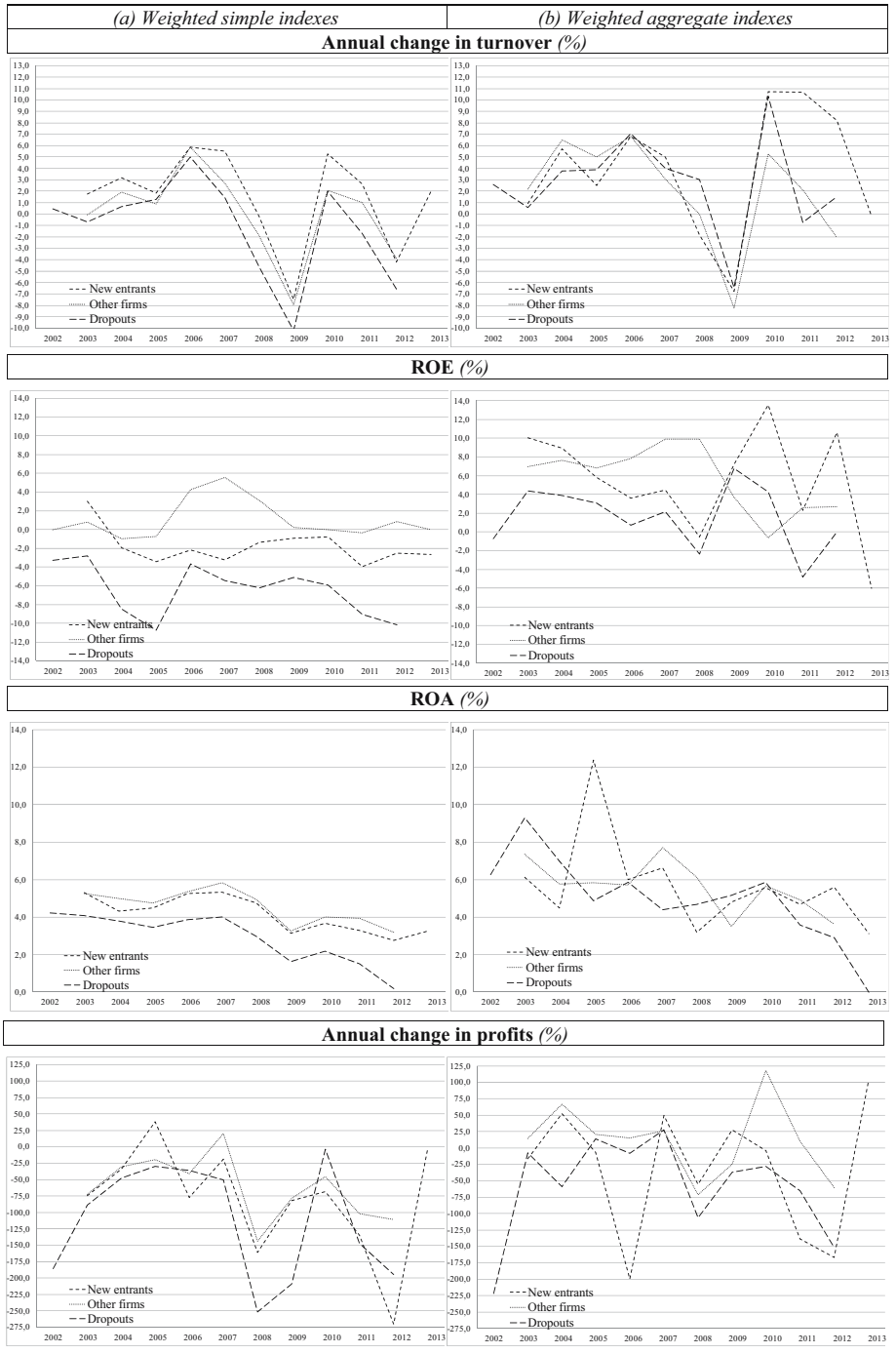

Fig. 2. Weighted indexes (percent relative changes) for new entrants in each survey edition, dropouts in the following edition and rest of the sample. Extreme values for individual values trimmed at the 1st and 99th percentiles. Firms affected by mergers, acquisitions and splits not considered. Source: Bank of Italy's Invind business survey and Cerved archive.

Table 5. Invind survey- OLS with the economic indicators as dependent variables Coefficients of the dummy for new entrants against next dropouts (2003–2012)[(a)(b)(c)(d)].

	Units considered: new entrants and future dropouts in each survey wave											
	Firms with 20 employee and over				*Only firms with 20–99 employees*				*Only firms with 100 employees and over*			
	Dependent variables				Dependent variables				Dependent variables			
	Change in turnover	ROE	ROA	Change in profit	Change in turnover	ROE	ROA	Change in profit	Change in turnover	ROE	ROA	Change in profit
2003	4.754***	11.646***	1.756***	11.717	5.602***	13.787***	1.465**	2.260	3.370**	7.612	2.486***	17.304
2004	3.516***	8.539***	0.698	8.964	3.523**	11.721***	1.010	18.118	3.714**	1.592	−0.161	−10.509
2005	5.032***	7.253***	1.144***	49.654	5.446***	7.939**	1.213**	134.435*	3.889**	6.196	0.969	−132.132
2006	2.842***	3.144	1.578***	−89.084	4.028***	3.304	1.665***	−75.658	−0.107	3.008	1.187	−126.148
2007	4.273***	8.320***	1.639***	17.152	5.400***	5.942	1.324**	52.456	2.147	12.851**	2.334***	−38.408
2008	4.432***	4.206*	2.097***	91.924	4.543***	5.059	1.962***	157.863	4.491	2.752	2.543***	−0.776
2009	2.857**	8.326***	1.916***	169.874**	3.401**	9.018***	1.887***	135.254	1.117	6.858	1.802*	218.863*
2010	5.806***	3.040	1.364***	−7.883	7.297***	4.930	1.318**	16.961	2.977	−0.500	1.613*	−60.449
2011	4.916***	9.376***	1.872***	106.387	7.004***	8.271**	2.003***	29.780	1.467	12.133***	1.763**	271.930**
2012	4.041***	13.570***	3.807***	29.590	4.000**	14.757***	4.103***	97.645	4.355**	11.148**	3.159***	−127.287
Pooled estimates[(e)]	**4.280*** **	**6.211*** **	**1.859*** **	**38.830****	**4.900*** **	**7.036*** **	**2.007*** **	**62.800*** **	**3.200*****	**4.979*** **	**1.624*** **	**0.010**
	Average share of new entrants and future dropouts over the total sample											
New entrants (A)	**17.1**				**21.9**				**18.0**			
Future dropouts (B)	**9.2**				**13.9**				**11.9**			
(A) + (B)	**26.3**				**35.8**				**29.9**			

(a) Dependent variables are expressed in percent. The table shows the coefficients and the significance of the dummy for new entrants into the sample, with future dropouts from the sample as reference. – (b) Separate OLS estimates. – (c) ***: p-value less than 0.01, **: p-value between 0.01 and 0.05, *: p-value between 0.05 and 0.1. Standard errors estimated with White's correction for heteroskedasticity. – (d) Additional covariates: sector of economic activities, location of firm's head office, class size. – (e) Standard errors are computed by considering the same firm repeated over time as a cluster. Additional covariates: dummies for the survey years.

Source: Bank of Italy's Invind business survey and Cerved archive.

account the time effect, as well as dummies for the survey design variables, that also control for the possible nonignorability of the survey design.

6.2.1. Modeling the Propensity to Enter a Wave

We estimate the probability to enter a wave relative to the year t. The firms modeled are those with at least 20 employees in the year $t - 1$ in the *Cerved* archive, provided they did not participate in the wave $t - 1$. Our dependent is a binary variable, indicating whether a firm will be in the wave relative to year t. The economic indicators used as covariates (we also use per capita turnover and profits) are modeled one at a time. Since each indicator is relative to the year $t - 1$, before an eligible unit enters the sample or still remains outside, it contributes to causally explain the propensity to enter the wave.

The standard error of the coefficients accounts both for the repeated observations relative to the same firm in the pooled data set and for the variability generated by using imputed values for the missing number of employees. The number of employees is required because it is a model covariate and also because it is needed to identify whether a firm belongs to the reference population of firms with 20 employees and over. The standard error of each coefficient β_k can therefore be written as:

$$SE(\beta_k) = \sqrt{V_M(\beta_k) + V_I(\beta_k)} \tag{5}$$

where $V_M(\beta_k)$ is the usual variance of the *logit* model coefficients, after correcting for the clustering effect derived by repeatedly using the same firm (Rogers, 1993) and $V_I(\beta_k)$ is the variance of the coefficient estimates computed over ten independent replicates of the imputation process.

The coefficients for ROA, ROE, profits change and per capita profits are positive and significant (Table 6). For turnover, the evidence is mixed, since the coefficient for per capita turnover is positive and not significant, whereas the one relative to turnover change is negative but weakly significant.

The overall result is that favorable economic indicators tend to positively associate with the propensity to enter a wave, with the strongest significance measured for ROA and ROE. The sign and the significance of the coefficients relative to the dummies for firm size (Table S4 in Supplementary material) indicate that, for the firms with less than 200 employees, the propensity to enter the sample is lower than for the firms with at least 500 employees. A geographical effect also emerges, because new entrances are more easily found among firms with head office in the two macro-areas of Italy - Center and Italy - South and isles, compared with those headquartered in the macro-area of Italy - North. This effect is also due to the sampling rate for firms headquartered in the macro-area of Italy - North structurally lower than that of the rest of the sample.

A relevant question is whether the performance gap between new entrants and rest of the reference population persists, once the latter ones enter the sample. We evaluate the issue by estimating the same *logit*, with the only difference that the economic indicators refer to the year when the new entrants enter the sample, instead of the previous one (Table 7). The positive relationship between the propensity to enter a wave and positive economic performances weakens, since the sizes of the positive coefficients for ROE and ROA decrease; those for profit change and per capita profit also decrease to the point of

Table 6. Invind survey: logit for the propensity of firm's entrance into the next wave (2002–2012)[(a)].

	(1)	(2)	(3)	(4)	(5)	(6)
Economic indicators						
roa	0.01060***					
roe		0.00185***				
change in profit per employees (%)			0.00274**			
change in profit (%)				0.00305*		
change in turnover per employee (%)					0.00134	
change in turnover (%)						−0.03680*
Number of observations	954,792	924,484	538,836	669,012	562,645	693,179

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.

(a)***: p-value less than 0.01, **: p-value between 0.01 and 0.05, * = p-value between than 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster and by accounting for the variability of the model used to impute the number of employees when missing. All estimates include dummies for the years.

Source: Bank of Italy's Invind business survey and Cerved archive.

Table 7. Invind survey: logit for the relationship between firm's entrance into a wave and firm's characteristics (2003–2013)[(a)].

	(1)	(2)	(3)	(4)	(5)	(6)
Economic indicators						
roa	0.00729***					
roe		0.00075***				
change in profit per employees (%)			0.00014			
change in profit (%)				0.00102		
change in turnover per employee (%)					−0.27500***	
change in turnover (%)						−0.11980***
Number of observations	845,013	784,121	578,360	821,253	604,420	849,515

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.

(a)***: p-value less than 0.01, **: p-value between 0.01 and 0.05, * = p-value between than 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster and by accounting for the variability of the model used to impute the number of employees when missing. All estimates include dummies for the years.

Source: Bank of Italy's Invind business survey and Cerved archive.

becoming not significant; as for turnover, the coefficient relative to the change goes negative, whereas the one of per-employee turnover remains negative and increases in size (Table S5 in Supplementary material reports the complete details).

How to explain this result? The quick search for firms that replace those unwilling to further participate is a common need for all panel business surveys, since the interviewers face strict time constraints. *Invind* is not an exception. The firms with above-average economic performances in the most recent years are naturally easier to enrol, since their managers can devote more time to fill the complex survey questionnaire, but they later tend to be more similar to comparable firms, as sometimes measured in the economic literature of firm behavior (Knapp et al. 2006).

6.2.2. Modeling the Propensity to Drop Out of a Wave

In order to model the propensity to drop out of a wave, we consider all its units. Following the same steps as in the propensity to be a new entrant, the dependent for the *logit* is a binary variable indicating whether a firm in wave t also participates in the following wave $t + 1$. The economic indicators used as covariates are relative to the year t, so that they can explain firm behavior at time $t + 1$. In the analysis, the number of employees is never missing, since it is collected in the survey, therefore the standard errors of the coefficients must only be corrected for the repeated observations of the same firm, derived from using the pooled data set.

The results (Table 8) are a mirror of those relative to the propensity to enter the sample: negative economic performances in a given year augment the propensity to leave the sample in the following year. The coefficients with the greatest sizes are those of ROA and ROE. Survey participation is more erratic for the firms with less than 200 employees (the effect is even greater for those with less than 50 employees, as seen in Table S6 in Supplementary material). A geographical effect is also present, since the enterprises headquartered in the the macro-area of Italy - Northeast tend to exit the survey more frequently than those belonging to the reference macro-area of Italy - South and isles. The greater size of the subsample of the firms headquartered in the Northeast, compared to those of the other areas, helps contain the risk of excessively reducing the sample size for this area.

7. The Effects of the Delays in Updating the Reference Population

The population distribution is required to compute the survey weights, which are necessary for obtaining sample estimates representative for the reference population. The official distribution of Italian enterprises is available with a two-year lag. This means that the population of the year $t - 2$ is used for the first estimates, released at the end of May of the year $t + 1$, relating to the survey for year t, carried out in the first months of $t + 1$. In the following years, the survey weights are re-computed as soon as the updated population distributions become available.

The delay can have relevant effects on the estimates because of their structure. As shown in Section 3, the survey relating to year t collects the values for the main interest variables for the years $t - 1$, t, and $t + 1$, so that relative changes $(t/t - 1)$ and $(t + 1/t)$ are computed by using data only from this wave. The procedure would require three distinct reference populations for the three years, but the approach is not followed, since

Table 8. Invind survey: logit for the propensity of firm's exit from the next wave (2002–2012)[(a)].

	(1)	(2)	(3)	(4)	(5)	(6)
Roa	−0.02440***					
roe		−0.00390***				
change in profit per employees (%)			−0.00501***			
change in profit (%)				−0.00504***		
change in turnover per employee (%)					−0.24800***	
change in turnover (%)						−0.44670***
Number of observations	38,708	37,732	34,828	36,027	41,275	41,279

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.

(a) ***: p-value less than 0.01, **: p-value between 0.01 and 0.05, * = p-value between than 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster. All estimates include dummies at level of year.

Source: Bank of Italy's Invind business survey and Cerved archive.

the survey weights would have to be repeatedly updated over the years. For practical reasons, only one set of survey weights is used for each survey edition, relative to the year t, which are updated as soon as the correct population information becomes available, with a very limited change in the original estimates.

Looking at the trend in the survey reference population, its size remained stable until 2006 (Figure 3), but deep demographic changes took place in the following years, since the population size increased in 2007–2008. A slow and steady decrease subsequently occurred during the years 2009–2012 of the economic crisis. The downsizing of the nonconstruction industrial sector in the years 2001–2012 was partially offset by the growing number of nonfinancial service-sector firms. This fact is largely due to the structural changes of the Italian economy, which led to a decrease in the share of GDP produced by the industrial sector. To a lesser degree, it also led to the changes of the classification criteria of economic activities, which shifted a range of activity (such as product maintenance and customer support) previously classified as industrial to the services sectors (Istat 2010).

Given these population changes, it is important to measure the error caused by using a unique set of weights. Our analysis focuses on turnover, employment levels and investment collected in the survey and most widely used. First of all, we assess the aggregate indexes of change, since they are extensively commented in the form of percent variation (Bank of Italy 2017, 66).

We express the indexes as percent changes. If we use a single weighting system, the weighted aggregate index can be written as: $\left(\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} - 1\right) 100$, if two separate weighting sets for the two periods $t-1$ and t are available, it becomes: $\left(\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t-1}} - 1\right) 100$.

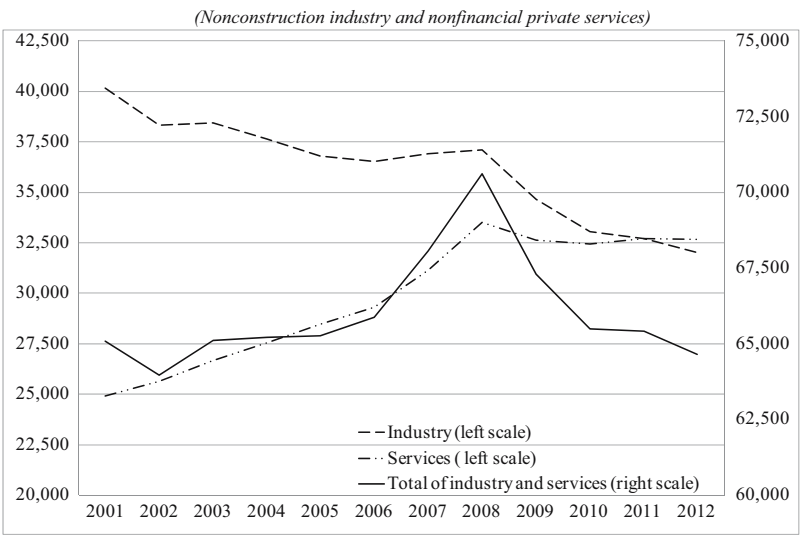

Fig. 3. Number of Italian firms with 20 employees and over, 2001–2012. Source: Istat, Italian Statistical Business Register (ASIA).

It is quite easy to verify that the weighted aggregate index with the single set of weights is smaller than the one with two sets of weights whenever the population size increases, whereas it is bigger if the population becomes smaller (see last page of Supplementary material for details).

Looking at the population trends (Figure 3) and their effects on these estimates (Figure 4, upper panel), the downward bias caused by the single weighting system was strong for the years 2007–2009 (especially for investment change in 2008), when the biggest population increase occurred. In the following years, the upward bias was small, because the decrease rate of the population size was steady, but rather limited. The greatest bias shows up for the relative change of employment, as expected, since it is the estimate most affected by repeated updates of the population distribution.

We finally deal with the effect of a unique weighting system on the simple indexes, since applied econometricians use individual changes derived from the survey in their microeconometric models. These models can also be estimated in a weighted version (Cameron and Trivedi 2005, 817), in order to produce estimates valid for the whole reference population. In such a case, it is relevant to assess the effect produced on the average of these individual changes by a weighting system not adequately representative of such a population. We focus our attention on the simple indexes of the changes of turnover and employment and we disregard investment changes, which are very erratic and unsuitable for microeconometric modeling (Doms and Dunne 1998).

We again express the indexes as percent changes. If we use a single weighting system, the weighted simple index can be written as: $\left(\frac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} w_{i,t} - 1\right) 100$. With two distinct weighting sets for $t-1$ and t, it can be expressed as:

$$\left(\frac{1}{\sum_{i=1}^{n_t} w_{i,t-1\cap t}} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} w_{i,t-1\cap t)} - 1\right) 100 \tag{6}$$

Here, $w_{i,t-1\cap t}$ is the weight referred to the firms belonging to the reference population both at time $t-1$ and at time t. We compute these weights by considering the population size $N_{h,t-1\cap t}$ of every stratum h as the minimum between those of the periods:

$$N_{h,t-1\cap t} = min\{N_{i,t-1}, N_{i,t}\} \tag{7}$$

This expression is an upper bound of the true value, with a negligible approximation error if the number of entrances and exits in the population are small relative to the population size in the two periods. For the reference population of *Invind*, the yearly balances of entrances and exits over the population are worth, on average, 4% for the strata and 5% for the poststrata considered in the survey design (the corresponding median values are 1.1% and 1.7%).

For turnover and employment changes, the double weighting system has a negligible effect on the simple indexes (Figure 4, lower panel).

An explanation of this result can be found by writing the difference between the same index computed respectively with the single and the double weighting system. For the

(*Invind*, 2002–2012)

(Extreme values trimmed to 1st and 99th percentiles)

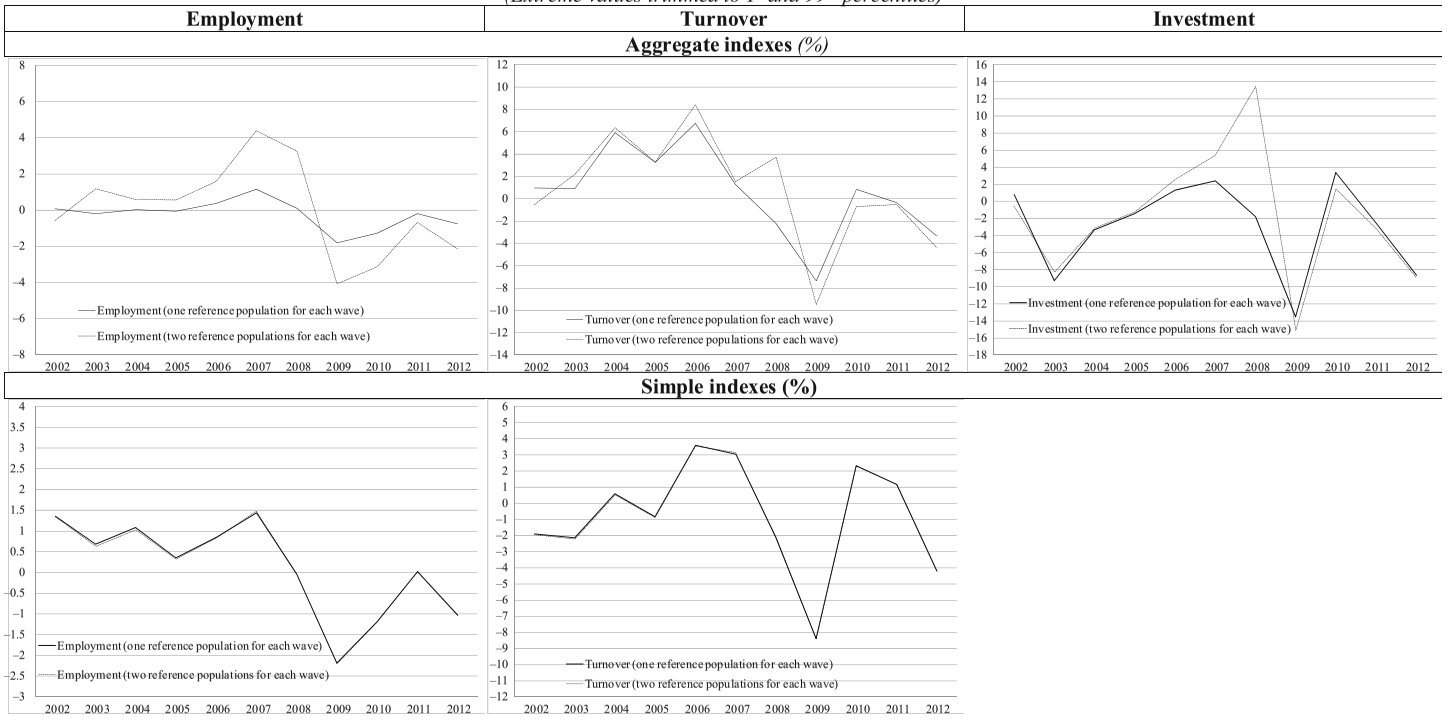

Fig. 4. Weighted indexes (percent relative changes) of the annual relative change of employment, turnover and investment Indexes expressed as per cent changes. Turnover and investment measured at constant 2013 prices. Source: Bank of Italy's Invind business survey.

weighted simple index, this expression is:

$$\sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \left(\frac{w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t}} - \frac{w_{i,t-1\cap t}}{\sum_{i=1}^{n_t} w_{i,t-1\cap t}} \right) \tag{8}$$

It is a sum of n_t terms, each formed by the product of two parts: the first one $\frac{y_{i,t}}{y_{i,t-1}}$ is always positive, the second part is small, since it is the difference between the ratios of the shares of single weights over corresponding weight totals, which tends to produce very low values for Equation 8.

Looking at the weighted aggregate index, the difference becomes:

$$\sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \left(\frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t} y_{i,t-1}} - \frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t-1} y_{i,t-1}} \right) \tag{9}$$

The structures of (8) and (9) are similar: they sum n_t terms, each formed by the product of two parts. The first one is $\frac{y_{i,t}}{y_{i,t-1}}$, always positive, the same in both expressions. The second part within round brackets is not necessarily limited to small values in (9) and the size of the expression accordingly tends to be greater than that of Equation 8.

8. Conclusions

Our article evaluates the sampling bias of a panel business survey by using auxiliary information derived from administrative data. We consider the Bank of Italy's *Invind* survey, for which two external sources are available:

- an archive of firms' financial statements, providing complete information for all the years 2002–2013 considered for every unit participating in the survey, regardless of the continuity of its participation;
- aggregate data on the sample reference population, available for the years between 2002 and 2012.

We use the first source to evaluate the bias caused by panel attrition on the indexes of yearly changes of variables such as turnover and profits and on averages of composite indicators like ROA and ROE. We focus on the attrition caused by firms leaving the panel in a given survey edition, replaced by others in the following one.

We find the estimates that are strongly dependent on big firms' values are less affected by panel attrition than the estimates representing the average behavior of firms that do not take firm size into account.

Looking more closely at the attrition determinants, positive economic performances make it easier to enroll new firms in the survey, in order to replace firms dropping out because of negative economic performances. However, the economic results of new entrances become more aligned to those of the population, once they enter the sample.

The statistical literature (Deville et al. 1993; Särndal and Lundström 2005) proposes many adjustments to the survey weights, some of them specifically designed for business surveys (Lavallée and Labelle-Blanchet 2013). In the form of generalized post-stratification procedures, these adjustments could be useful in compensating for the bias caused by panel attrition (see also Faiella 2010 and Solon et al. 2015, for a comprehensive

review of the cases when the utilization of survey weights is advisable). For example, these techniques could be studied in future research to produce new survey weights that incorporate the information on sample firms' past economic performances.

Our analysis also suggests some possible revisions of the survey management First of all, more efforts could be devoted to prevent firms with negative economic results from dropping out of the sample at the current rate. With this aim, better interviewer training could highlight the importance of maintaining the most difficult units in the sample.

We finally analyze the effects of using survey weigths computed with a population distribution that is not up-to-date. The issue is relevant, since a single *Invind* edition collects data on the values of the interest variables for three consecutive years $t - 1$, t and $t + 1$ ($t + 1$ is a forecast that refers to the year when the interviews are conducted). The current solution is to use only the information relative to the reference population at time t, which is subsequently updated. We show how this solution entails the risk of bias for the weighted aggregate indexes of relative changes of turnover, investment and employment levels. This bias tends to be negative when the population size increases, but it becomes positive when the population size decreases, without substantial effects on the simple indexes, which do not depend on firm size. This risk can be avoided by computing, for every survey editions relative to the year t, separate sets of weights for the different years considered. This is feasible for older survey editions, while, for the most recent ones, external evidence on the population trends should be carefully used to assess the quality of the estimates.

9. References

Afonso, L.M. 2015. "Correcting for Attrition in Panel Data Using Inverse Probability Weighting: An application To the EU15 Bank System." *Doctoral dissertation (Lisbon School of Economics and Management, Working Paper)*. Available at: https://www.repository.utl.pt/bitstream/10400.5/8155/1/DM-LMA-2015.pdf (accessed January 2019).

Ardilly, P. and P. Lavallée. 2007. "Weighting in Rotating Samples: The SILC survey in France." *Survey Methodology* 33(2): 131–137.

Bank of Italy. 2005. *Supplements to the Statistical Bulletin, Sample Surveys, Survey of Industrial and Service firms*, Year 2003, Volume XV, 20 October 2005. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2003-indagini-imprese/en_suppl_55_05.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2014. *Supplements to the Statistical Bulletin, Sample Surveys, Survey of Industrial and Service firms*, Year 2013, New Series, Year XXIV, 24 July 2014. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2013-indagine-imprese/en_suppl_40_2014.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2015. *Supplements to the Statistical Bulletin, Sample Surveys Survey of Industrial and Service firms*, Year 2014, New Series, Year XXV, 1st July 2015. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2014-indagine-imprese/en_suppl_34_2015.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2017. *Annual Report for 2016*. Available at: https://www.bancaditalia.it/pubblicazioni/relazione-annuale/2016/index.html (accessed January 2019).

Black, C., D.C. Broadstock, A. Collins, and L. Hunt. 2007. "A Practical Guide to Developments in Data Imputation Methods." *Traffic Engineering and Control* 48(8): 358–363. Available at: https://www.seec.surrey.ac.uk/research/Publications/BlackBroadstockCollins&Hunt(2007).pdf (accessed january 2019).

Cameron, A.C. and P.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

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

Deng, Y., D.S. Hillygus, J.P. Reiter, Y. Si, and S. Zhen. 2013. "Handling Attrition in Longitudinal Studies: The Case for Refreshment Samples." *Statistical Science* 28(2): 238–256. Doi: http://dx.doi.org/10.1214/13-STS414.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88(423): 1013–1020. Doi: http://dx.doi.org/10.1080/01621459.1993.10476369.

Doms, M. and T. Dunne. 1998. "Capital Adjustment Patterns in Manufacturing Plants." *Review of Economic Dynamics* 1(2): 409–429. Available at: http://www.homepages.ucl.ac.uk/~uctpjrt/doms%26dunne.pdf (accessed January 2019).

Fabbris, L. 1989. *L'indagine campionaria*. Nuova Italia Scientifica.

Faiella, I. 2010. "The use of survey weights in regression analysis." *Bank of Italy's Working Paper* n. 739. Doi: http://dx.doi.org/10.2139/ssrn.1601936.

Hollander, M. and D.A. Wolfe. 1973. *Nonparametric Statistical Methods*. John New York: Wiley.

Istat. 2010. Available at: http://www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007 (accessed January 2019).

Istat. 2014. *Archivio Asia*. Avaiable at: http://www.istat.it/it/archivio/archivio+asia (accessed January 2019).

Knapp, M., A. Gart, and M. Chaudhry. 2006. "The Impact of Mean Reversion of Bank Profitability on Post-merger Performance in the Banking Industry." *Journal of Banking and Finance* 30(12): 3503–3517.

Lavallée, P. and S. Labelle-Blanchet. 2013. "Indirect Sampling Applied to Skewed Populations." *Survey Methodology* 39(1): 183–215.

Little, R. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Martin, E., D. Abreu, and F. Winters. 2001. "Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation." *Journal of Official Statistics* 17(2): 267–284. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/money-and-motive-effects-of-incentives-on-panel-attrition-in-the-survey-of-income-and-program-participation.pdf (accessed January 2019).

Rogers, W. 1994. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 3(13). Available at: https://EconPapers.repec.org/RePEc:tsj:stbull:y:1994:v:3:i:13:sg17 (accessed January 2019).

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley.

Solon, G., S.J. Haider, and J. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* (2): 301–316. Doi: http://dx.doi.org/10.3368/jhr.50.2.301.

Trivellato, U. 1999. "Issues in the Design and Analysis of Panel Studies: a Cursory Review." *Quality & Quantity* (33): 339–352. Doi: http://dx.doi.org/10.1023/A:1004657006031.

Received July 2016
Revised January 2018
Accepted April 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 93–115, http://dx.doi.org/10.2478/JOS-2019-0005

The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias

Barbara Felderer[1], *Antje Kirchner*[2], *and Frauke Kreuter*[3]

More and more surveys are conducted online. While web surveys are generally cheaper and tend to have lower measurement error in comparison to other survey modes, especially for sensitive questions, potential advantages might be offset by larger nonresponse bias. This article compares the data quality in a web survey administration to another common mode of survey administration, the telephone.

The unique feature of this study is the availability of administrative records for all sampled individuals in combination with a random assignment of survey mode. This specific design allows us to investigate and compare potential bias in survey statistics due to 1) nonresponse error, 2) measurement error, and 3) combined bias of these two error sources and hence, an overall assessment of data quality for two common modes of survey administration, telephone and web.

Our results show that overall mean estimates on the web are more biased compared to the telephone mode. Nonresponse and measurement bias tend to reinforce each other in both modes, with nonresponse bias being somewhat more pronounced in the web mode. While measurement error bias tends to be smaller in the web survey implementation, interestingly, our results also show that the web does not consistently outperform the telephone mode for sensitive questions.

Key words: Mode effects; telephone survey; web survey; combined bias.

1. Introduction

Researchers often use evidence of bias in survey estimates to assess and compare data quality among different modes of survey administration. There are two major problems with this approach. First, by assessing combined bias as a measure of data quality, researchers mix different sources of bias, for example, bias due to differential coverage, nonresponse or measurement error which might each differ in magnitude and direction, and do so differently for different survey modes. In the worst case, seemingly unbiased estimates in one survey mode might actually be more biased compared to another survey mode, when each source of bias is investigated individually. Hence, investigating combined bias leaves researchers guessing about the sources of bias and makes it hard to

[1] University of Mannheim, Mannheim, 68131, Germany. Email: felderer@uni-mannheim.de
[2] RTI International, 3040 E. Cornwallis Road, RTP, NC 27709, U.S.A. Email: akirchner@rti.org
[3] University of Maryland, College Park, Maryland 20742, U.S.A. Email: jkreuter@umd.edu

Acknowledgments: This material is partly based upon work supported by the National Science Foundation under Grant No. SES-1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Stephanie Eckman, Julie Korbmacher, Renae Reis, and Joe Sakshaug for helpful comments on earlier drafts of this article.

derive practical implications and inform survey designs. Understanding the individual contributions of, for example, nonresponse and measurement error bias to the total survey error and potential bias, and how these differ by survey mode is of utmost importance (Biemer 2010). Another common challenge that researchers face is how to actually measure bias. Often, researchers compare sample estimates to other aggregate population estimates, or they rely on assumptions, such as the "more-is-better assumption" for undesirable behaviors, to assess which survey mode performs better. This approach does not necessarily inform researchers about which mode is the least biased and comes closest to the "truth." Ideally, bias is assessed by comparing survey estimates with auxiliary and gold standard data for the same sampled individuals. This measure can then be used to inform the research community about the overall effects of survey mode on data quality. However, often researchers do not have access to this kind of information; either because the data are nonexistent, for example when investigating attitudes, or, the data are unavailable for reasons of data confidentiality.

The key features of this study that allow us to address both challenges are its experimental design and the use of a unique combination of large scale survey data and administrative records from German social security records – containing rich information on a variety of labor market related and demographic characteristics. The specific design of our study enables us not only to measure bias directly but also to disentangle different sources of error contributing to bias among two commonly used survey modes, telephone and web. More specifically, we separate bias due to nonresponse and measurement error, which ultimately helps researchers to understand the nature and relative contribution of each bias source to combined bias. These results are particularly relevant for researchers planning to use a mixedmode design, in which it is generally assumed that the strengths of one mode will compensate the weaknesses of another, and thus enhance data quality, while at the same time potentially reducing costs.

The following section will provide a brief overview of why differences in bias between survey modes are to be expected, how bias has been assessed in past studies, and the research questions of our article. Section 2 will introduce the design, data and methods used in our analyses. The results are described in Section 3. The article concludes with a summary and discussion of the main results in Section 4.

1.1. Why Do We Expect Differences in Bias Across Survey Modes?

Both data collection modes, telephone and web, have their particular strengths and weaknesses with respect to achieving survey participation and response accuracy. Survey mode can affect the sample composition, as different modes have different coverage error. In order to participate in a telephone survey, sample members have to have access to a telephone, whereas for web surveys, access to the Internet is a prerequisite. In list-assisted sampling designs this might lead to differential coverage error if for example, more sample members have access to a telephone than the Internet. To the extent that coverage error is systematic and differs by mode, this might introduce differential bias. Sample composition may also differ, as the ability to establish contact with the target person and respondents' willingness and capacity to complete the survey differs across survey modes (Dillman et al. 2002, 6). Self-administered surveys, for

example, tend to have lower response rates compared to interviewer-administered surveys (De Leeuw 2005). Additionally, if relevant subgroups self-select depending on the survey mode, this can introduce differential nonresponse bias, should the selection mechanism be related to survey variables of interest (Groves and Couper 1998; De Leeuw et al. 2008; Biemer 2010; Kreuter et al. 2010). Measurement error results from a difference in the respondents' survey report and the (unobserved) true value, for example, due to misunderstanding a question, failure to retrieve the correct information or incorrect reporting (for a review, see, for example Biemer and Lyberg 2003). If the misreporting mechanism is related to the survey outcome of interest and systematically differs between survey modes, this again will result in differentially biased estimates.

In addition to coverage, nonresponse and measurement error, there are other potential sources of error that may bias survey estimates, including specification or adjustment errors (Biemer 2010). In line with existing research, we will focus on bias due to nonresponse and measurement error when investigating mode differences, as those can be expected to be the main drivers of differential bias. Previous empirical research on mode differences shows that response rates tend to be generally lower in web surveys (Lozar Manfreda et al. 2008). This increases the potential for selectivity and *nonresponse bias* in the web survey compared to the telephone survey (Fricker et al. 2005; O'Neill and Dixon 2005; Abraham et al. 2006; Groves 2006; Letourneau and Zbikowski 2008). Although response rates are known to be lower in web surveys, web surveys have several benefits over more traditional surveying methods. Web surveys are generally less cost intensive, the data are available almost immediately, respondents can take the survey at their own pace and convenience, and it provides a more private survey setting (Callegaro et al. 2014). Due to this latter fact, *measurement error bias* might be less pronounced for certain types of questions in the self-administered web mode compared to interviewer-administered modes (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Sakshaug et al. 2010). While we would expect to see little difference between web and telephone for factual items that are less prone to misreporting (Atkeson et al. 2014), survey mode might influence response accuracy for items that are sensitive in nature or those that evoke concerns of social (un)desirability (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Sakshaug et al. 2010; Atkeson et al. 2014).

In line with the literature on sensitive questions (Lee 1993; Groves 2004; Bradburn et al. 2004), traits that are positively valued – such as regular employment – should be overreported, while undesirable traits – such as welfare receipt or marginal employment – should be underreported. While a respondent might choose to give a correct answer to a sensitive question in the web mode due to the increased privacy, they might respond differently to the same question in a telephone interview with an interviewer present (Kreuter et al. 2008; Chang and Krosnick 2009, 2010; Malhotra et al. 2014; Roberts et al. 2014). Hence overall, the combined bias due to nonresponse and measurement error might actually be smaller in the web administration compared to the telephone mode. Again, if one were to investigate combined bias only, researchers would never know how the bias terms interact. The main focus in our article will be the interaction of nonresponse and measurement error and how each contributes to bias. We will discuss other potential sources of bias as appropriate.

1.2. How has Bias been Assessed in the Past?

Mode effects studies are usually not able to directly differentiate nonresponse bias and measurement error bias in a survey estimate, but instead rely on indirect indicators, benchmarks or assumptions (such as "more-is-better" for sensitive questions) for a comparison of data quality in the absence of gold standard validation data. External population benchmarks for some variables are often used to assess nonresponse bias (Fricker et al. 2005; Yeager et al. 2011; Malhotra et al. 2014). Bias due to measurement error is often assessed using indirect indicators of survey satisficing, including non-differentiation, item missingness, the use of extreme values, acquiescence or socially desirable responses (McCabe et al. 2002; Duffy et al. 2005; Chang and Krosnick 2009, 2010; Atkeson et al. 2011, 2014; Hope et al. 2014; Malhotra et al. 2014). Other mode effects studies use panel information from previous waves to assess bias due to nonresponse and measurement error (Sax et al. 2003; Duffy et al. 2005; Braunsberger et al. 2007; Chang and Krosnick 2009; Vannieuwenhuyze et al. 2010; Roberts et al. 2014). The results of these studies are not comparable with cross-sectional studies, since nonresponse or measurement error bias for the initial wave is usually not assessed and data provided in the initial wave need not necessarily be more accurate. Also, typically none of these studies analyze the combined effect of nonresponse and measurement error on survey estimates.

The most powerful designs to study differential effects of survey mode on nonresponse bias and measurement error bias are those that, in addition to random assignment of survey mode, have auxiliary data with extraordinary data quality available for all sample cases for the characteristics under study. Few studies allow for a validation study comparing web surveys to other forms of survey administration (e.g., McCabe et al. 2002; Sax et al. 2003; Sanders et al. 2007; Kreuter et al. 2008; Dillman et al. 2009; Sakshaug et al. 2010; Atkeson et al. 2011; Stephenson and Crête 2011; Atkeson et al. 2014). Particularly relevant for our study are the results of the validation studies by Kreuter et al. (2008) and Sakshaug et al. (2010), as they focus on the interaction of both sources of bias for a variety of variables and question types. Kreuter et al. (2008) find significant differences in completion rates comparing telephone, interactive voice recording, and web. The initial screening interview was conducted by phone, and screener respondents assigned to the web had the lowest completion rates. However, the results do suggest that sensitive items are reported more accurately in the web mode. Regarding the interaction of both error sources, both studies find that bias due to measurement error dominates nonresponse error for sensitive items, while nonresponse error tends to be larger for neutral and socially desirable items (Kreuter et al. 2008; Sakshaug et al. 2010). For the most part, Sakshaug et al. (2010) find that the different error sources reinforce each other and do not cancel each other out.

The main contribution of our article is a systematic assessment of the relative contribution of nonresponse and measurement error to the combined bias in survey estimates in each survey mode using large scale validation data that is known to be of very good data quality and can serve as a gold standard. Building on past validation studies by Kreuter et al. (2008) and Sakshaug et al. (2010), we analyze the interaction of nonresponse and measurement error for demographic and sensitive questions. Whereas these studies focus on a very specific subpopulation of student alumni, the scope of our analyses is

broader. Our analyses rely on a stratified random sample of the adult labor-force in Germany. As such, more generalizable inferences can be drawn regarding the implications of the choice of a particular survey mode. Existing validation studies also typically analyze nonresponse or measurement error bias for mean statistics of certain survey items, but do not assess bias in distributions. Our analysis also investigates and compares bias in distributions for two metric items – age and income.

2. Data and Methods

2.1. Study Design and Administrative Data

The Integrated Employment Biographies (IAB 2011) maintained by the German Federal Employment Agency (FEA) serve as the sampling frame for our study. This administrative register combines information on individuals' times of (un)employment in Germany and welfare, also called basic income support ("Unemployment Benefit II", abbreviated UB II). These registers cover approximately 86% of the German labor force, starting from 1975, including all employees who are subject to social security contributions, individuals seeking employment, and those on welfare, excluding only self-employed and civil servants. This sampling frame is comprehensive, up to date (with only a short time lag) and accurate, since it contains payment-relevant information, as their main use is by the German statutory pension insurance to administer and calculate pension claims, benefit claims, and payments. For the analysis, we use updated versions of the data sets that are used to generate the IEB (IAB 2012, 2013).

Sampling from the FEA registers provides us with detailed information on (un)employment and welfare benefit receipt for all sampled cases, that is, respondents and nonrespondents to the survey. Sampled individuals were randomly assigned to one of the two modes. We specifically designed and worded both surveys such that survey responses can be validated given the information in the administrative data, including socio-demographic (e.g., gender and age) and sensitive information (e.g., income and welfare benefit receipt). Furthermore, we only use data that are known to be accurate and complete, and can thus serve as gold standard (Jacobebbinghaus and Seth 2007).

More specifically, we investigate nonresponse bias, measurement error bias, and combined bias in gender (0 male, 1 female), mean age (and categories: 18–29, 30–39, 40–49, 50–59, 60+ years), currently employed (0 no, 1 yes), type of employment (marginal employment with an income of EUR 400 and less, regular employment with an income of EUR 401 and more), past receipt of UB II (past 12 months), and mean monthly labor income (and income terciles) in euros, if currently employed. In Germany, respondents think in terms of monthly income and not annual income. Thus, the survey items ask for monthly income. However, labor income in the administrative records is captured only as the total gross income in a given employment spell (typically one year). Thus, monthly income has to be derived from this measure. The basic assumption is that all income is equally distributed over the months of a certain spell. Also, income is top coded in the administrative data, the limit being a yearly income of approximately EUR 57,000 in the states of former East Germany and EUR 66,000 in the states of former West Germany, depending on the type of pension insurance. Since this affects all

administrative data equally and survey mode was randomly assigned, inferences with respect to the relative comparison across modes are still valid.

Administrative data used for the analyses were extracted from either the last valid (employment) period or, in the case of an ongoing period, from the respective interview month for respondents. The date of the last interview in either mode is taken as the reference date for nonrespondents.

2.2. *Survey Data*

Overall, a sample of 24,236 eligible adults was drawn in June 2011 from the FEA registers and randomly assigned to one of two survey modes: 12,400 individuals were randomly assigned to the telephone mode, while 11,836 individuals were assigned to complete the survey online. Addresses, and in part telephone numbers, were available for all sampled individuals in the frame.

Only 9,332 of the individuals assigned to the telephone mode turned out to have valid phone numbers. 2,400 individuals completed the telephone survey, corresponding to an overall response rate of 19.35% among sample members in the telephone mode of the experiment (RR1 according to AAPOR 2011). In the web mode, 1,311 letters were returned to sender due to an incorrect address, leaving 10,525 individuals who received the invitation to the survey. Of those, 1,082 individuals completed the web survey. The overall response rate among sample members in the web survey was 9.14%. Table 1 provides an overview of the sample sizes and response rates.

The telephone survey was fielded during the months of August to October 2011 and the web survey from February to mid-April 2012. Prior to fieldwork, all sampled individuals received an advance/invitation letter inviting them to participate in the government survey "Work and Consumption in Germany", commissioned by the Institute for Employment Research (IAB), and carried out by the LINK Institute. The invitation letter for the web mode also contained all relevant login information and a conditional incentive of EUR 3. Two weeks after the start of fieldwork, a reminder was sent to all sampled cases of the web survey component.

Both questionnaires contained questions relating to employment biographies that are conceptually equivalent to the administrative data described above. We only analyze questions that were fielded in exactly the same way in both surveys, except for one question about past receipt of unemployment benefit and will provide more information below. Both surveys were kept as similar as possible and only differed in some of the

Table 1. Response rates across modes of data collection.

	Telephone survey	Web survey
Sampled	12,400	11,836
Valid contact information	9,332	10,525
Completed	2,400	1,082
Response rate (ref. sampled)	19.35%	9.14%

questions in other parts of the questionnaire. The average survey completion time was 21 minutes for the telephone interview and 15 minutes in the web mode.

2.3. Methods

In order to assess nonresponse bias, we only include individuals for whom we have valid contact information (see Table 1) as all other individuals never received the invitation to participate in the survey. For the telephone, this means that we include individuals for whom we have a valid telephone number. Given that we do not have this kind of information for individuals assigned to the web mode, we include those who actually received our invitation to participate in the survey, that is, individuals whose invitation letter was not returned to sender. This approach implies that while we can clearly separate bias due to coverage error and nonresponse for individuals assigned to the telephone; the same does not hold for those individuals assigned to the web mode. A small portion of the nonresponse bias that we investigate will actually be coverage bias, although we expect this to be minimal as the internet penetration in Germany is quite high. Approximately 79% of the German households had internet at the time of the survey, with an additional 14% of the noninternet households having internet access outside home, for example, at work (Statistisches Bundesamt 2013). Furthermore, as we are comparing survey packages, our results give a realistic assessment of relative biases in these two survey modes. For simplicity we will refer to this as nonresponse bias in both survey modes. In a sensitivity analysis, we replicated our analysis including all sample cases, for example, all individuals assigned to the telephone mode including those without valid telephone numbers and all individuals assigned to the web mode including those whose invitation letter was returned to sender. Results can be found in Subsection 5.2. Appendix B.

Respondents to both modes are part of the measurement error analysis. Due to data protection regulations we are not able to match the data on an individual level, but rather compare the proportion of respondents reporting a certain characteristic in the survey with the proportion of respondents who have this same characteristic in the administrative data. Because we analyze survey and administrative data separately and do not combine data sources, we are not restricted to those respondents who consented to data linkage, and can include all survey respondents. While this has the advantage that our study is not subject to potential linkage nonconsent bias (Sakshaug and Kreuter 2012), it has the disadvantage that we cannot examine measurement error at an individual level.

We compare bias in mean statistics (and distributions) for the variables introduced above across both modes using the:

a) full sample administrative data (fs): $\frac{1}{N}\sum_{i=1}^{N} y_{i,\text{admin}}$ with N being the sample size;
b) respondent sample administrative data (resp): $\frac{1}{n}\sum_{i=1}^{n} y_{i,\text{admin}}$ with n being the number of completed interviews; and
c) respondent sample survey data (svy): $\frac{1}{n}\sum_{i=1}^{n} y_{i,\text{survey}}$.

For an assessment of the combined bias we will compare a) the true value from the full sample administrative data to c) the respondent sample survey data. We will then break this combined bias into its components: nonresponse bias is assessed by comparing

estimates from the full sample administrative data (a) to estimates from respondent sample administrative data (b). Bias due to measurement error is assessed by comparing estimates based on the respondent sample administrative data (b) to estimates based on respondent sample survey data (c).

As information is available for all sample cases, the estimation of nonresponse bias is straightforward. Nonresponse bias (nr) is the difference of the true nonrandom sample value according to administrative records (adm) for the full sample (fs) and the mean computed using the respondents (resp) only. In order to compare nonresponse bias between variables and modes, nonresponse bias is standardized by the full sample mean of each variable of interest multiplied by 100 to obtain the relative nonresponse bias in percent:

$$\widehat{\boldsymbol{rel.bias}}(\hat{\bar{y}}_{nr}) = \frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100. \tag{1}$$

As our analysis focuses on relative biases in mean statistics and not on the mean statistics themselves, we need to estimate the variances of the relative biases and adapt significance tests accordingly.

More specifically, the variance of $\widehat{\boldsymbol{rel.bias}}(\hat{\bar{y}}_{nr})$ is given by:

$$Var(\widehat{\boldsymbol{rel.bias}}(\hat{\bar{y}}_{nr})) = Var\left(\frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100\right). \tag{2}$$

As $\bar{y}_{adm,fs}$ is nonrandom, we can write:

$$\begin{aligned} Var(\widehat{\boldsymbol{rel.bias}}(\hat{\bar{y}}_{nr})) = \frac{100^2}{\bar{y}^2_{adm,fs}} (&Var(\hat{\bar{y}}_{adm,resp}) + Var(\bar{y}_{adm,fs}) \\ &+ 2Cov(\bar{y}_{adm,fs}, \hat{\bar{y}}_{adm,resp})). \end{aligned} \tag{3}$$

From $\bar{y}_{adm,fs}$ being nonrandom, it also follows that $Var(\bar{y}_{adm,fs}) = 0$ and $Cov(\bar{y}_{adm,fs}, \hat{\bar{y}}_{adm,resp}) = 0$. Thus, the variance reduces to:

$$Var(\widehat{\boldsymbol{rel.bias}}(\hat{\bar{y}}_{nr})) = \frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{adm,resp}). \tag{4}$$

This leads to the test statistic for a one-sample Z-Test for evaluating the significance of individual relative biases:

$$z = \frac{\frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100}{\sqrt{\frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{adm,resp})}}. \tag{5}$$

Under the null hypothesis $Var(y_{adm,resp})$ equals $Var(y_{adm,fs})$ so we can substitute $Var(\hat{\bar{y}}_{adm,resp})$ by $Var(y_{adm,fs})/n$ with n denoting the respondent sample size:

$$z = \frac{\frac{\hat{\bar{y}}_{adm,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}}}{\sqrt{\frac{Var(y_{adm,fs})}{\bar{y}^2_{adm,fs} n}}} \tag{6}$$

The two-sample Z-Test for comparing the relative biases in the telephone and web mode is then given as:

$$z = \frac{\left(\frac{\hat{\bar{y}}_{adm,resp,web} - \bar{y}_{adm,fs,web}}{\bar{y}_{adm,fs,web}}\right) * 100 - \left(\frac{\hat{\bar{y}}_{adm,resp,cati} - \bar{y}_{adm,fs,cati}}{\bar{y}_{adm,fs,cati}}\right) * 100}{\sqrt{\frac{100^2}{\bar{y}^2_{adm,fs,web}} Var(\hat{\bar{y}}_{adm,resp,web}) + \frac{100^2}{\bar{y}^2_{adm,fs,cati}} Var(\hat{\bar{y}}_{adm,resp,cati})}}. \tag{7}$$

Transforming the counter and substituting $Var(\hat{\bar{y}}_{adm,resp,cati})$ by $Var(y_{adm,fs,web})/n_{web}$ and $Var(\hat{\bar{y}}_{adm,resp,cati})$ by $Var(y_{adm,fs,cati})/n_{cati}$ we derive:

$$z = \frac{\left(\frac{\hat{\bar{y}}_{adm,resp,web}}{\bar{y}_{adm,fs,web}} - \frac{\hat{\bar{y}}_{adm,resp,cati}}{\bar{y}_{adm,fs,cati}}\right)}{\sqrt{\frac{Var(y_{adm,resp,web})}{\bar{y}^2_{adm,fs,web} n_{web}} + \frac{Var(y_{adm,resp,cati})}{\bar{y}^2_{adm,fs,cati} n_{cati}}}}. \tag{8}$$

Similar to the estimation of nonresponse bias, bias due to measurement error (me) is straightforward to calculate, as the true values are known from the administrative records. Bias due to measurement error is given as the difference of the mean estimate in the survey data (svy) and the true statistic according to administrative records for all respondents. Standardizing measurement error bias with the mean of the respondents based on the administrative data multiplied by 100 gives us an estimate of the relative bias in percent:

$$\widehat{rel.bias}(\hat{\bar{y}}_{me}) = \frac{\hat{\bar{y}}_{svy,resp} - \bar{y}_{adm,resp}}{\bar{y}_{adm,resp}} * 100. \tag{9}$$

In the comparison of respondent sample survey data and respondent sample administrative data, the respondent sample administrative data are taken as the nonrandom gold standard, as they contain the true information of the full sample of respondents. This implies that for this analysis we assume $Var(\bar{y}_{adm,resp} = 0)$ and $Cov(\hat{\bar{y}}_{svy,resp}, \hat{\bar{y}}_{admin,resp}) = 0$ leading to:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{me})) = \frac{100^2}{\bar{y}^2_{adm,resp}} Var(\hat{\bar{y}}_{svy,resp}). \tag{10}$$

Like all survey data, some of the survey items are subject to item nonresponse. Very few respondents do not report an employment status (telephone 0.2%; web 3.3%) and past receipt of UB II (telephone 0.2%; web 2.8%). Since we are estimating the proportion of respondents belonging to a certain employment or past UB II status (yes/no), missing information is implicitly treated as a "no" response (e.g., not employed, no UB II receipt) in the assessment of measurement bias. The proportion of item nonresponse is highest in the income information (telephone 13.8%; web 15.7%) which results in a reduction of the

case base for the survey estimates which is used in the measurement error analysis. There is no item nonresponse in the reports of gender or age. In a sensitivity analysis, we drop cases with missing information in employment and past UB II status for the corresponding analysis. Neither of our results reported below change substantively.

The combined bias (combined) due to nonresponse and measurement error for a survey statistic is simply the difference between the estimate derived from the full sample administrative data and the respondent sample survey data. The combined bias estimate can be standardized similarly to the other biases to obtain the relative combined bias in percent.

$$\widehat{rel.bias}(\hat{\bar{y}}_{comb}) = \frac{\hat{\bar{y}}_{svy,resp} - \bar{y}_{adm,fs}}{\bar{y}_{adm,fs}} * 100. \tag{11}$$

Comparing respondent sample survey data and full sample administrative data, the full sample administrative data means are nonrandom, implying $Var(\bar{y}_{adm,fs} = 0)$ and $Cov(\hat{\bar{y}}_{svy,resp}, \bar{y}_{admin,fs}) = 0$ leading to:

$$Var(\widehat{rel.bias}(\hat{\bar{y}}_{comb})) = \frac{100^2}{\bar{y}^2_{adm,fs}} Var(\hat{\bar{y}}_{svy,resp}). \tag{12}$$

Test statistics for relative measurement error bias and combined bias can be derived in an identical manner as for relative nonresponse bias in Equation 5 and Equation 8.

For the subsequent analyses, we distinguish demographic information (such as gender and age), from potentially sensitive information (type of employment, past receipt of UB II and mean labor income from current employment) and report the results in that order. To reiterate our expectations: irrespective of question type, we would expect there to be a generally lower nonresponse bias in the telephone mode compared to the web. We expect little to no measurement error bias for demographic items in either mode, whereas sensitive questions should be reported more accurately in the web mode. The prediction for combined bias depends on whether both sources of bias enforce or compensate each other.

3. Results

We report the results for each error source by indicator and only report differences in the text that are statistically significant at an alpha level of 0.05 ($p < 0.05$), based on the adapted Z-Tests. Figures 1 to 4 display the relative bias in mean estimates for different variables separated by horizontal lines, by survey mode each due to nonresponse (nr), measurement error (me) and combined bias (combined), including 95%-confidence intervals. Solid triangles indicate bias for the telephone mode and hollow squares indicate bias for the web mode. The dashed vertical line indicates zero percent relative bias. Relative bias estimates, including confidence intervals and test statistics can be found in Subsection 5.1. Appendix A.

Females are significantly overrepresented in both survey modes with biases significantly differing between the two modes (see Figure 1). Not surprisingly, there is virtually no measurement error bias for gender in either mode. The very small discrepancies might be due to the fact that some individuals might identify with a gender other than the sex originally recorded in the administrative data. An overrepresentation of

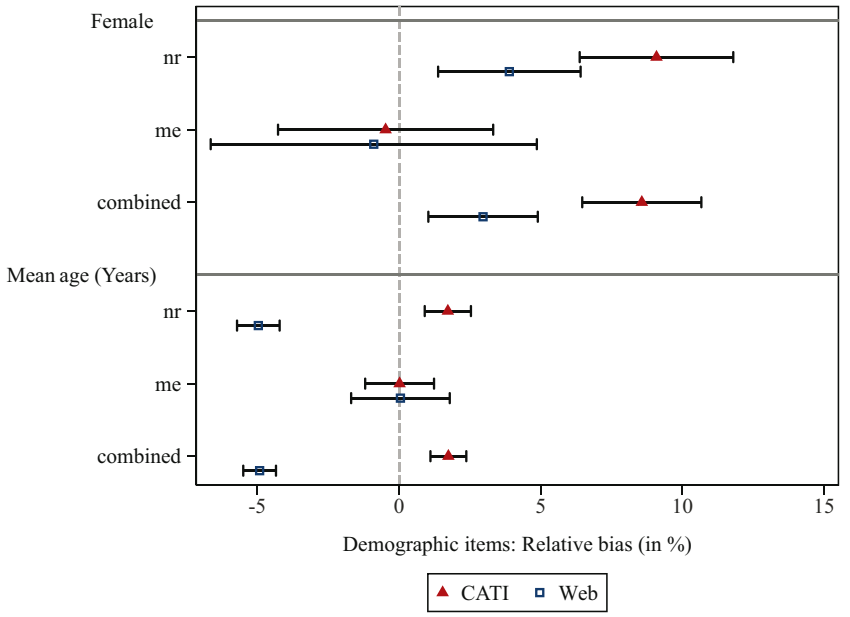

Fig. 1. Relative combined bias for socio-demographic variables, including 95%-confidence intervals.

females together with a very small measurement error bias leads to a relative combined bias that is dominated by nonresponse bias, that is, a significant overestimation of the proportion of women in both survey modes. The relative combined bias is significantly higher in the telephone mode than the web mode.

Our results also show a significant negative nonresponse bias for mean *age* in the web mode and a significant positive nonresponse bias in the telephone mode, although smaller in magnitude. Substantively, this means that younger individuals are overrepresented in the web mode, whereas older individuals are overrepresented in the telephone mode. The difference in biases between the two modes is statistically significant. As expected, there is virtually no relative measurement error bias in mean age for either mode. Combined bias is therefore almost identical to nonresponse bias and implies a significant overestimation of mean age in the telephone mode and significant underestimation in the web mode compared to the population. Relative combined bias differs significantly between the two survey modes.

To study bias in age in more detail, we investigate biases in several *age categories* (see Figure 2). In line with our expectations, younger individuals are overrepresented in the web mode while middle-aged and older individuals are overrepresented in the telephone mode, although relative nonresponse bias is not always significantly different from zero. Except for the middle-aged category "aged 40–49 years" biases differ significantly between both modes. Similarly to mean age, there is no evidence for significant measurement error in any of the age categories, with relative measurement error biases being very close to zero. Again, this results in a combined bias that is almost identical in magnitude to that of nonresponse bias: the proportion of younger individuals is overestimated in the web survey, whereas the proportion of individuals in the middle-aged and older age categories are overestimated relying on telephone survey estimates.

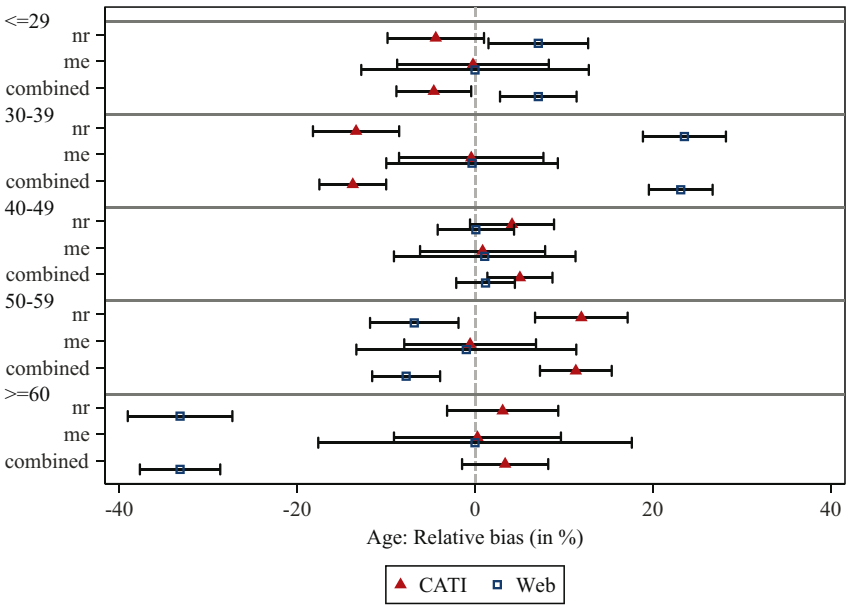

Fig. 2. Relative combined bias for age distribution, including 95%-confidence intervals.

Although combined bias is not significantly different from zero for every mode and age category, relative combined bias differs significantly between survey modes for all age categories except "age 40–49."

We now turn to those items potentially subject to social desirability concerns and sensitivity displayed in Figure 3. Our results suggest that relative nonresponse bias in *employment status* points in the same direction for both modes such that employed

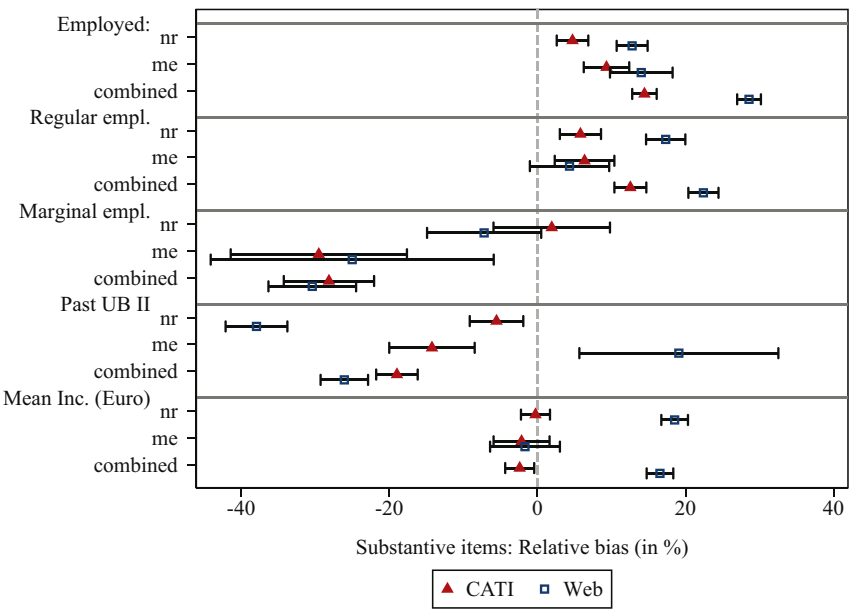

Fig. 3. Relative combined bias for substantive variables, including 95%-confidence intervals.

individuals are significantly overrepresented. This overrepresentation is significant for both survey modes and is significantly higher in the web mode than in the telephone mode. Investigating the different types of employment, we see that individuals in a regular form of employment are significantly overrepresented in both modes. This overrepresentation is, again, significantly higher for web mode than the telephone mode. There is no evidence of significant nonresponse bias among marginally employed individuals. Relative nonresponse bias in the employment variables tends to be larger in the web mode compared to the telephone mode.

Turning to bias due to measurement error, in line with our theoretical expectations, the socially more desirable characteristic of regular employment is significantly overreported in the telephone mode, but does not show significant measurement error bias in the self-administered web mode. However, regular employment is only slightly more accurately estimated in the web mode than in the telephone mode, with the difference in biases not being statistically significant. The potentially more stigmatizing form of marginal employment is significantly underreported in both modes, although to a somewhat lesser extent in the web mode. Like for regular employment, biases do not differ significantly between the modes. Again, we attribute these results to social desirability: telling an interviewer that one has a regular job is more desirable and less of a norm violation than admitting to being "only" marginally employed. Hence, not surprisingly, relative bias due to measurement error is always slightly higher in the telephone mode compared to the web mode although these differences are not significant for any employment type across modes. With the exception of marginal employment in the telephone mode, relative nonresponse and relative measurement bias reinforce each other, leading to an even larger relative combined bias. Despite a marginally smaller measurement error bias, the web mode exhibits a consistently larger combined bias compared to the telephone mode (differences are not statistically significant for marginal employment).

Relative nonresponse bias in *past benefit receipt* is negative for both modes. This leads to a significant underestimation of the proportion of individuals who received welfare in the past year and this underestimation is significantly more pronounced in the web mode than in the telephone mode. Relative measurement error bias points in different directions for both modes such that the proportion of past benefit recipients is overestimated in the web mode and underestimated in the telephone mode (with differences being statistically significant). Surprisingly, the magnitude of measurement bias is strikingly similar across both modes. Relative nonresponse and measurement bias reinforce each other in the telephone mode and point in opposite directions in the web mode. Despite this compensation of both sources of bias in the web mode, relative combined bias is still significantly larger compared to the telephone mode.

Mean *income* is significantly biased due to nonresponse in the web mode, whereas there is no significant relative nonresponse bias in mean income in the telephone mode. The difference in relative nonresponse bias is significant. There is no significant relative measurement bias for mean income in either mode. Relative combined bias is statistically significant for both modes and is mostly driven by nonresponse bias. Whereas there is only a small negative combined bias in the telephone mode, this bias is much larger in the web mode, which results in an overestimation of mean income. We find that relative nonresponse and combined bias differ significantly between the two survey modes.

There is no significant bias due to nonresponse in the telephone mode in the *income categories* except for a slight overestimation of the lower income category. This differs in the web mode: individuals with a low income are significantly underrepresented, whereas those with a high income are significantly overrepresented. Relative nonresponse biases differ significantly between the modes. Although measurement error bias for mean income is statistically nonsignificant, both modes show considerable measurement error bias in the different income categories (Figure 4). While significantly more respondents claim to belong to the low income group (telephone) or the middle income group (web), in both modes too few respondents report that they belong to the highest income category. Measurement error bias does not differ significantly between both modes for any of the income categories. All income categories show significant combined bias for both modes pointing in opposite directions (and being significantly different between the modes) in the lowest and the highest income category. Combined bias is mostly driven by nonresponse bias in the web mode and measurement error bias in the telephone mode.

To summarize our results, the individual contributions of nonresponse and measurement error bias for those variables that show significant relative combined bias indicate that nonresponse bias exceeds measurement error bias in magnitude for gender in both modes, for mean age in both modes (and all categories except for 40–49 years in the web mode and ages older than 60 in the telephone mode), and mean income (as well as low and high income) in the web mode. On the other hand, measurement error bias is larger than nonresponse bias for employment, "marginal" employment in both modes and income (as well as income categories) in the telephone mode. Relative combined bias in regular employment and past unemployment benefit receipt differs in its composition across the modes: nonresponse bias is larger than measurement error bias for both characteristics in the web mode, whereas measurement error bias exceeds nonresponse bias in the telephone mode.

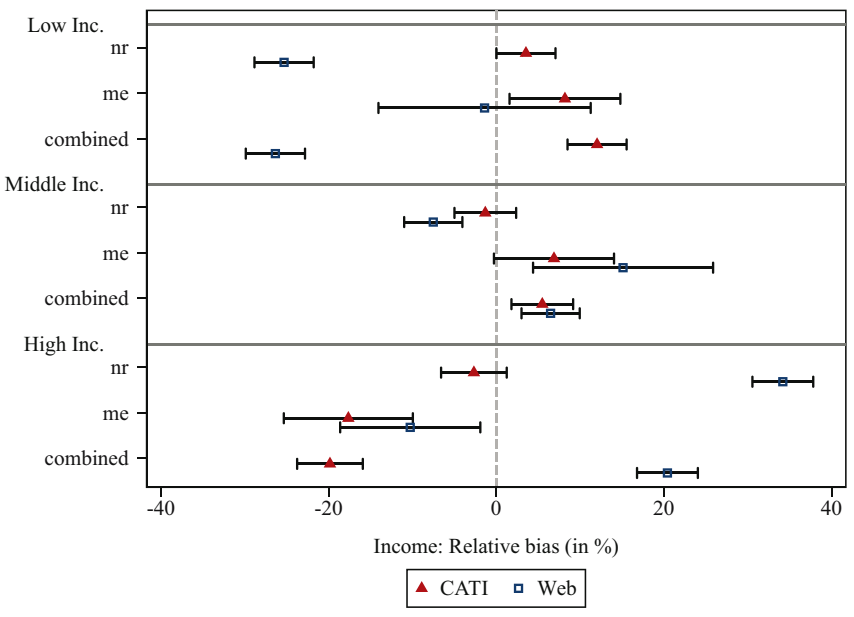

Fig. 4. Relative combined bias for income distribution, including 95%-confidence intervals.

With respect to the *interaction of the two sources of bias,* the results show that bias due to nonresponse and measurement error tend to reinforce each other, with the exception of past benefit receipt and the income categories in the web survey. Our results suggest that the relative combined bias is larger for the web mode compared to the telephone mode for mean age, employment status and employment type, past UB II receipt, mean labor income and the income categories. The combined relative bias is larger in the telephone mode than in the web mode for gender, whereas there is no consistent pattern across age groups. These results suggest that the data obtained via the web survey administration are, overall, more biased compared to the telephone mode.

4. Summary and Discussion

Our results show that the estimates obtained from the web survey are biased to a larger extent compared to the telephone survey when considering combined bias. In line with previous research, these results are mostly driven by a larger nonresponse bias in both modes for demographic items and by larger measurement error bias for sensitive items. Our results further suggest that potential social desirability concerns from respondents are somewhat alleviated in the web mode. However, the potential benefits of a smaller measurement error bias in the web mode are inconsistent across estimates and do not outweigh the comparatively larger nonresponse bias compared to the telephone mode.

The result for overreporting of past welfare benefit, UB II, receipt in the web mode is also somewhat puzzling. One potential explanation for the overreporting of welfare receipt could be slight differences in the question wording between the modes. The telephone survey asked for "welfare receipt in 2010", that is, the previous year, while the web survey asked for "welfare receipt in the past 12 months". The data collection period of the web survey was in the beginning of 2012, so we expect that some respondents did not refer to the last 12 months in their retrieval, but instead included the period since January 2011, thus leading to an overestimate. The administrative data can exactly differentiate these differing periods. Also, web survey respondents might have skipped reading the information regarding the reference period altogether (about 7% of those respondents who did not receive UB II in the reference period actually received benefits at some earlier point). Because respondents in the telephone component of the study received a series of filter questions about different earnings in 2011 and the respondents in the web mode only saw this one question, respondents in the web mode might be more prone to suffer from referring to the wrong reference period. Both kinds of error would result in web survey respondents reporting receipt prior to the 12-month reference period – since February 2011 – and thus explain the significant amount of overreporting. Another explanation for the underreporting in the telephone mode could also be due to a strategy to avoid follow-up questions. However, Eckman et al. (2014) find no significant filter effect for the income questions in this survey. These potential errors confound our results with respect to mode differences in UB II receipt. Nonetheless, web seems to outperform telephone for this item, in the sense that it is able to alleviate social desirability concerns.

We find substantial misreporting of income across different income categories, although this does not differ across modes. This is surprising in that we would have expected more accurate reporting in the web mode due to increased privacy and the fact that an individual

can take the survey at their own pace, potentially spending more time to retrieve accurate information. In line with results reported in previous studies (e.g., Duncan and Hill 1985; Bound and Krueger 1991; Rodgers et al. 1993), measurement error seems to be correlated with true income; more specifically, that it is mean-reverting, which is a tendency for those with lower earnings to overstate these and those with higher earnings to understate. The tendency to overreport the middle category to the disadvantage of the extreme categories can clearly be seen for the web mode.

Our results are subject to some limitations. First of all, the question arises as to whether nonresponse adjustment techniques would alleviate bias in an identical manner in both modes and how this would affect the combined bias. Since nonresponse bias tends to point in opposite directions for both modes, the most obvious solution might be to pool the samples (De Rada and del Amo 2014). Another option is to rely on different weighting techniques (Bethlehem 2010). However, such techniques are only reducing bias if weighting variables are correlated with both nonresponse and survey variables of interest (Kreuter and Olson 2011). Studying nonresponse bias before adjustment is a topic in its own right, as Schouten et al. (2016) conclude that balanced samples are always advantageous, regardless of adjustment techniques that might be applied in retrospect. However, if the same mechanisms that lead to nonresponse bias are related to measurement error bias (Olson 2013; Malhotra et al. 2014; Roberts et al. 2014), combined bias might actually be inflated. The comparison of both modes after these adjustments are particularly interesting, that is, whether bias estimates are affected in a similar manner. While that is an interesting research question, these analyses are beyond the scope of this article. The second limitation is that we are comparing respondents across mode "packages" and cannot directly attribute measurement differences to mode effects. For example, differences in responses might not be causally attributed to different reporting schemes evoked by different modes, but can also be driven by different individuals (with differential reporting behavior) responding to different modes (and hence be due to sample composition). Third, data collection periods in both modes differed slightly and there is a potential time effect that we cannot rule out. However, with the exception of one characteristic – past receipt of welfare – we are confident that this does not jeopardize our results.

To reiterate, our results are in line with previous research: while younger, employed and more educated individuals participate in the web survey, there is less bias in the telephone mode and nonresponse bias tends to point in opposite directions in both modes. At the same time, measurement error bias tends to be equivalent or smaller in the web mode. Given that the web mode has several advantages over the telephone mode with respect to survey costs and immediate data availability, one implication that follows from our results could be to implement a sequential mixed mode design, especially since nonresponse biases in both modes tend to be in opposing directions. Thus, approaching respondents first by web and then following up on nonrespondents by the telephone seems to be a promising approach to reach different subgroups in the population and balance the respondent sample. Another promising strategy to reduce measurement error bias in the telephone could be to supplement the telephone component with a self-administration mode, either using IVR, T-ACASI or a web add-on, each with its own advantages and disadvantages.

5. Appendix

5.1. Appendix A – Biases

Table 2. Relative nonresponse bias, relative measurement bias and relative combined bias including 95% confidence intervals. Significant biases, based on one-sample Z-tests are indicated in boldface (p0.05). Z-values are reported for two-sample Z-Test of differences between the telephone and web mode.

	Rel. nonresponse bias			Rel. measurement error bias			Rel. combined bias		
Variable	CATI	Web	z-value	CATI	Web	z-value	CATI	Web	z-value
Female	**9.08**	**3.89**	**− 2.75**	−0.48	−0.89	−0.12	**8.57**	**2.96**	**− 3.85**
	(6.37; 11.80)	(1.38; 6.41)		(−4.27; 3.32)	(−6.65; 4.86)		(6.46; 10.67)	(1.03;4.9)	
Mean age (years)	**1.72**	**− 4.97**	**− 11.82**	0.02	0.05	0.03	**1.74**	**− 4.92**	**− 15.24**
	(0.91; 2.54)	(−5.72; −4.22)		(−1.20; 1.23)	(−1.69; 1.79)		(1.11; 2.37)	(−5.50; −4.34)	
Age ≤ 29	−4.42	**7.12**	**2.90**	−0.23	0.00	0.03	**− 4.64**	**7.12**	**3.83**
	(−9.84; 1.01)	(1.51; 12.72)		(−8.76; 8.30)	(−12.79; 12.79)		(−8.84; −0.43)	(2.81; 11.42)	
Age 30–39	**− 13.38**	**23.55**	**10.76**	−0.43	−0.33	0.01	**− 13.75**	**23.14**	**13.92**
	(−18.23; −8.53)	(18.88; 28.22)		(−8.55; 7.69)	(−9.98; 9.31)		(−17.51; −9.99)	(19.55; 26.72)	
Age 40–49	4.16	0.09	−1.66	0.85	1.09	0.04	**5.05**	1.18	−1.54
	(−0.56; 8.88)	(−4.20; 4.39)		(−6.18; 7.88)	(−9.12; 11.3)		(1.38; 8.71)	(−2.11; 4.48)	
Age 50–59	**11.96**	**− 6.83**	**− 5.12**	−0.55	−0.98	−0.06	**11.34**	**− 7.75**	**− 6.73**
	(6.75; 17.16)	(−11.81; −1.86)		(−7.96; 6.86)	(−13.35; 11.39)		(7.30; 15.37)	(−11.57; −3.93)	
Age ≥ 60	3.11	**− 33.17**	**− 8.28**	0.27	0.00	−0.03	3.39	**− 33.17**	**− 10.82**
	(−3.14; 9.36)	(−39.06; −27.29)		(−9.12; 9.66)	(−17.63; 17.63)		(−1.46; 8.23)	(−37.69; −28.65)	
Employed	**4.74**	**12.79**	**5.29**	**9.33**	**14.03**	1.76	**14.46**	**28.58**	**12.02**
	(2.61; 6.88)	(10.70; 14.88)		(6.26; 12.40)	(9.80; 18.25)		(12.80; 16.11)	(26.97; 30.18)	
Regular employment	**5.81**	**17.33**	**5.87**	**6.37**	4.34	−0.59	**12.55**	**22.42**	**6.52**
	(3.03; 8.59)	(14.68; 19.98)		(2.35; 10.40)	(−1.01; 9.69)		(10.40; 14.71)	(20.38; 24.45)	
Marginal employment	1.94	−7.19	−1.63	**− 29.51**	**− 25**	0.39	**− 28.14**	**− 30.39**	−0.52
	(−5.92; 9.80)	(−14.89; 0.52)		(−41.40; −17.61)	(−44.10; −5.90)		(−34.24; −22.05)	(−36.31; −24.47)	
Mean income (EUR)	−0.26	**18.53**	**13.85**	−2.13	−1.67	0.15	**− 2.39**	**16.55**	**13.96**
	(−2.22; 1.70)	(16.73; 20.33)		(−5.91; 1.64)	(−6.38; 3.04)		(−4.34; −0.43)	(14.75; 18.35)	
Low income	**3.54**	**− 25.31**	**− 11.33**	**8.20**	−1.39	−1.31	**12.03**	**− 26.34**	**− 15.07**
	(0.02; 7.07)	(−28.84; −21.77)		(1.59; 14.81)	(−14.05; 11.27)		(8.51; 15.55)	(−29.87; −22.81)	
Middle income	−1.30	**− 7.51**	**− 2.40**	6.89	**15.13**	1.25	**5.50**	**6.49**	0.38
	(−4.98; 2.38)	(−10.98; −4.04)		(−0.27; 14.05)	(4.39; 25.87)		(1.82; 9.18)	(3.02; 9.96)	
High income	−2.66	**34.19**	**13.53**	**− 17.64**	**− 10.26**	1.27	**− 19.83**	**20.42**	**14.78**
	(−6.58; 1.25)	(30.56; 37.81)		(−25.34; −9.94)	(−18.62; −1.9)		(−23.75; −15.92)	(16.80; 24.05)	
Past receipt of UB II	**− 5.52**	**− 37.92**	**− 11.51**	**− 14.23**	**19.10**	**4.47**	**− 18.96**	**− 26.07**	**− 3.27**
	(−9.13; −1.90)	(−42.09; −33.75)		(−20.00; −8.46)	(5.67; 32.54)		(−21.76; −16.16)	(−29.27; −22.86)	

5.2. Appendix B – Biases when Using All Sample Cases

In a sensitivity analysis, we replicated the bias estimation, including all sample cases, for example, all individuals assigned to the telephone mode, including those without valid telephone numbers and all individuals assigned to the web mode, including those whose invitation letter was returned to sender (see Table 3). In this second analysis, bias due to deployability and coverage cannot be separated from nonresponse bias. For simplicity, we will continue to refer to this as nonresponse bias. While relative nonresponse and relative combined biases might be affected, relative measurement error biases stay the same as they only refer to the survey respondents.

Although relative nonresponse biases change in magnitude, the relative difference in a comparison of the survey modes does not change for any of the variables compared to the analysis excluding the individuals for whom we do not have valid contact information. We do find some significant differences: when including all cases, relative nonresponse bias for *mean income* is now significantly different from zero for both modes as opposed to the web mode only. Relative nonresponse bias in *29 years and younger* is not significantly different from zero in any mode and modes do not significantly differ from each other when including individuals without valid contact information, whereas relative nonresponse bias for this variable is significantly different from zero in the web and biases significantly differ between the modes when excluding individuals without valid contact information. Also, relative nonresponse bias in *ages 50–59 years* is not significant in the web mode when including individuals without valid contact information. The age group *60 years and older* shows significant relative nonresponse biases in both modes, with significantly different relative nonresponse biases between the modes when including individuals without valid contact information, whereas the relative nonresponse bias for the telephone survey loses significance when excluding those individuals. This results in individuals aged *29 years and younger* being overrepresented when including all cases, but underrepresented when excluding individuals without valid contact information for the telephone survey, although relative nonresponse bias is not significant for this age group in any analysis of the telephone survey. Strikingly, the negative effect of past welfare receipt turns from a negative to a larger positive effect in the telephone survey when including the individuals without valid contact information, with differences between the modes being significant in both kinds of analysis.

As for relative nonresponse bias, the magnitudes of the combined bias differ slightly for the two kinds of analyses, but the directionality and relative differences comparing the survey modes are not affected for most of the variables. Differences in the relative magnitude of combined bias can only be found for *age 29 years and younger* and *middle income*. The differences in relative combined bias between the survey modes are not significant for *middle income* when the individuals without valid contact information are excluded, but are significant if they are included. However, the relative bias in middle income is significantly different from zero for both modes in both kinds of analyses. For *age 29 years and younger* we find the difference in relative combined bias between the two modes to be significant when excluding the individuals without valid contact information, but to be not significant if including these individuals. This is mostly due to a change in directionality for the telephone mode and a shift towards zero for the web mode. The

Table 3. Relative nonresponse bias, relative measurement bias and relative combined bias, including 95% confidence intervals. Significant biases based on one-sample Z-tests are indicated in boldface ($p < 0.05$). Z-values are reported for two-sample Z-Tests of differences between the telephone and web mode.

	Rel. nonresponse bias			Rel. measurement error bias			Rel. combined bias		
Variable	CATI	Web	z-value	CATI	Web	z-value	CATI	Web	z-value
Female	**8.07** (5.77; 10.38)	**4.37** (1.96; 6.78)	**− 2.18**	−0.48 (−4.27; 3.32)	−0.89 (−6.65; 4.86)	−0.12	**7.56** (5.75; 9.37)	**3.44** (1.61; 5.26)	**− 3.14**
Mean age (years)	**1.27** (0.58; 1.96)	**− 4.07** (−4.79; −3.34)	**− 10.46**	0.02 (−1.20; 1.23)	0.05 (−1.69; 1.79)	0.03	**1.29** (0.75; 1.83)	**− 4.02** (−4.57; −3.47)	**− 13.53**
Age ≤ 29	4.91 (−0.02; 9.84)	2.11 (−3.09; 7.31)	−0.76	−0.23 (−8.76; 8.30)	0.00 (−12.79; 12.79)	0.03	**4.67** (0.80; 8.53)	2.11 (−1.82; 6.05)	−0.91
Age 30–39	**− 16.51** (−20.57; −12.44)	**18.72** (14.38; 23.06)	**11.60**	−0.43 (−8.55; 7.69)	−0.33 (−9.98; 9.31)	0.01	**− 16.86** (−20.05; −13.68)	**18.32** (15.03; 21.61)	**15.07**
Age 40–49	−1.60 (−5.51; 2.30)	2.14 (−2.01; 6.30)	1.29	0.85 (−6.18; 7.88)	1.09 (−9.12; 11.30)	0.04	−0.77 (−3.83; 2.29)	**3.26** (0.11; 6.41)	1.80
Age 50–59	**12.9** (8.40; 17.39)	−3.49 (−8.34; 1.37)	**− 4.85**	−0.55 (−7.96; 6.86)	−0.98 (−13.35; 11.39)	−0.06	**12.27** (8.75; 15.79)	**− 4.43** (−8.11; −0.76)	**− 6.43**
Age ≥ 60	**5.77** (0.32; 11.22)	**− 30.86** (−36.59; −25.13)	**− 9.08**	0.27 (−9.12; 9.66)	0.00 (−17.63; 17.63)	−0.03	**6.05** (1.79; 10.32)	**− 30.86** (−35.20; −26.52)	**− 11.89**
Employed	**2.67** (0.88; 4.45)	**16.10** (14.03; 18.16)	**9.65**	**9.33** (6.26; 12.40)	**14.03** (9.80; 18.25)	1.76	**12.18** (10.79; 13.58)	**32.33** (30.77; 33.89)	**18.84**
Regular employment	**2.51** (0.19; 4.83)	**20.55** (17.95; 23.15)	**10.16**	**6.37** (2.35; 10.40)	4.34 (−1.01; 9.69)	−0.59	**9.04** (7.23; 10.85)	**25.78** (23.81; 27.75)	**12.26**
Marginal employment	3.02 (−3.77; 9.81)	−4.27 (−11.75; 3.22)	−1.41	**− 29.51** (−41.4; −17.61)	**− 25.00** (−44.10; −5.90)	0.39	**− 27.38** (−32.70; −22.06)	**− 28.20** (−33.87; −22.53)	−0.21
Mean income (EUR)	**− 6.55** (−8.19; −4.91)	**18.91** (17.18; 20.63)	**20.97**	−2.13 (−5.91; 1.64)	−1.67 (−6.38; 3.04)	0.15	**− 8.54** (−10.18; −6.90)	**16.92** (15.20; 18.64)	**20.96**
Low income	**11.24** (8.04; 14.44)	**− 25.38** (−28.75; −22.01)	**− 15.44**	**8.2** (1.59; 14.81)	−1.39 (−14.05; 11.27)	−1.32	**20.36** (17.16; 23.55)	**− 26.41** (−29.79; −23.04)	**− 19.73**
Middle income	**3.36** (0.09; 6.63)	**− 8.88** (−12.16; −5.60)	**− 5.18**	6.89 (−0.27; 14.05)	**15.13** (4.39; 25.87)	1.25	**10.48** (7.21; 13.75)	**4.90** (1.63; 8.18)	**− 2.36**
High income	**− 13.76** (−16.83; −10.69)	**36.51** (33.00; 40.02)	**21.12**	**− 17.64** (−25.34; −9.94)	**− 10.26** (−18.62; −1.90)	1.27	**− 28.97** (−32.05; −25.90)	**22.51** (19.00; 26.02)	**21.63**
Past receipt of UB II	**9.21** (5.76; 12.66)	**− 38.19** (−42.17; −34.22)	**− 17.65**	**− 14.23** (−20.00; −8.46)	**19.10** (5.67; 32.54)	**4.47**	**− 6.33** (−9.03; −3.63)	**− 26.39** (−29.39; −23.38)	**− 9.72**

effects of *age 40–49 years* in the web and *60 years and older* in the telephone mode increase and are significant in this second analysis. For *age 40–49 years* the relative combined bias in the telephone survey is less pronounced and not significant when including the individuals without valid contact information.

Even though we find some differences for relative nonresponse bias and relative combined bias between the two analyses for some age and income categories, these differences do not substantively change our findings and do not affect relative combined bias in mean age or mean income. The only substantive difference between the two sets of analysis is the change from underrepresentation to overrepresentation of past recipients of welfare benefit in the telephone mode when including the individuals without valid contact information. From this, we can conclude that more valid telephone numbers have been available for past benefit recipients than for nonrecipients. This makes sense, as individuals on UB II have to provide the German Federal Employment Agency with their telephone numbers to manage benefit claims. Even though this affects the relative combined bias in our survey, we do not expect this to be a general finding as this is very specific to the sample drawn using the data from the German Federal Employment Agency.

6. References

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

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: 676–703. Doi: http://dx.doi.org/10.1093/poq/nfl037.

Atkeson, L.R., A.N. Adams, and M.R. Alvarez. 2014. "Nonresponse and Mode Effects in Self- and Interviewer-Administered Surveys." *Political Analysis* 22: 304–320. Doi: http://dx.doi.org/10.1093/pan/mpt049.

Atkeson, L.R., A.N. Adams, L.A. Bryant, L. Zilberman, and K.L. Saunders. 2011. "Considering Mixed Mode Surveys for Questions in Political Behavior: Using the Internet and Mail to Get Quality Data at Reasonable Costs." *Political Behavior* 33: 161–178. Doi: http://dx.doi.org/10.1007/s11109-010-9121-1.

Bethlehem, J. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78: 161–188. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00112.x.

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

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

Bound, J. and A.B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1–24. Doi: http://dx.doi.org/10.3386/w2885.

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

Braunsberger, K., H. Wybenga, and R. Gates. 2007. "A Comparison of Reliability Between Telephone and Web-Based Surveys." *Journal of Business Research* 60: 758–764. Doi: http://dx.doi.org/10.1016/j.jbusres.2007.02.015.

Callegaro, M., R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavrakas. 2014. *Online Panel Research. A Data Quality Perspective*. Chichester: Wiley.

Chang, L. and J.A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet. Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73: 641–678. Doi: http://dx.doi.org/10.1093/poq/nfp075.

Chang, L. and J.A. Krosnick. 2010. "Comparing Oral Interviewing With Self-Administered Computerized Questions: An Experiment." *Public Opinion Quarterly* 74: 154–167. Doi: http://dx.doi.org/10.1093/poq/nfp090.

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

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

De Rada, V.D. and S.P. del Amo. 2014. "Two Are Better Than One: The Use of a Mixed-Mode Data Collection to Improve the Electoral Forecast." *Survey Practice* 7: 1–6. Doi: http://dx.doi.org/10.29115/SP-2014-0003.

Dillman, D.A., J.L. Eltinge, R.M. Groves, and R.J.A. Little. 2002. "Survey Nonresponse in Design, Data Collection and Analysis." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 3–26. New York: Wiley.

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: 1–18. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.007.

Duffy, B., K. Smith, G. Terhanian, and J. Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47: 615–639. Doi: http://doi.org/10.1177/147078530504700602.

Duncan, G. and D. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data." *Journal of Labor Economics* 3: 508–532. Doi: http://dx.doi.org/10.1086/298067.

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

Fricker, S., M. Galesic, R. Tourangeau, and T. Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 6: 370–392. Doi: http://dx.doi.org/10.1093/poq/nfi027.

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

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. Doi: http://dx.doi.org/10.1093/poq/nfl033.

Groves, R.M. and M. Couper. 1998. Nonresponse in Household Interview Surveys. Wiley Series in Probability and Statistics: Survey Methodology Section. New York: Wiley.

Hope, S., P. Campanelli, G. Nicolaas, P. Lynn, and A. Jäckle. 2014. "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration." *ISER Working Paper Series* No. 2014-20: 1–41. Available at: http://hdl.handle.net/10419/123808 (accessed December 2014).

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2011. Nuremberg: Integrierte Erwerbsbiographien (IEB) V09.00.

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2012. Nuremberg: Leistungshistorik Grundsicherung (LHG), Version 06.06.

IAB (Institut für Arbeitsmarkt- und Berufsforschung). 2013. Nuremberg: Beschäftigtenhistorik (BeH), Version 09.03.00.

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

Kreuter, F. and K. Olson. 2011. "Multiple Auxiliary Variables in Nonresponse Adjustment." *Sociological Methods & Research* 40: 311–332. Doi: http://dx.doi.org/10.1177/0049124111400042.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. Ezatti-Rice, C. Casas-Cordero, A. Petychev, R. M. Groves, and T. Raghuatan. 2010. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173: 389–407. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00621.x.

Kreuter, F., S. Presser, and R. Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys. The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72: 847–865. Doi: http://dx.doi.org/10.1093/poq/nfn063.

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

Letourneau, P.M. and A.A. Zbikowski. 2008. "Nonresponse in the American Time Use Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 4, 2008. 1283–1290. Denver, CO: American Statistical Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2008/Files/300982.pdf (accessed April 2018).

Lozar Manfreda, K., M. Bosnjak, J. Berzelak, I. Haas, and V. Vehovar. 2008. "Web Surveys Versus Other Survey Modes. A Meta-Analysis Comparing Response Rates." *International Journal of Market Research* 50: 79–104. Doi: http://dx.doi.org/10.1177/147078530805000107.

Malhotra, N., J.M. Miller, and J. Wedeking. 2014. "The Relationship Between Nonresponse Strategies and Measurement Error. Comparing Online Panels to Traditional Surveys." In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J. Krosnick, and P.J. Lavrakas, 313–336. Chichester: Wiley.

McCabe, S.E., C.J. Boyd, M.P. Couper, S. Crawford, and H. D'Arcy. 2002. "Mode Effects for Collecting Alcohol and Other Drug Use Data: Web and U.S. Mail." *Journal of Studies on Alcohol* 63: 755–761. Doi: http://dx.doi.org/10.15288/jsa.2002.63.755.

Olson, K. 2013. "Do Non-Response Follow-Ups Improve or Reduce Data Quality? A Review of the Existing Literature." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 176: 129–145. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01042.x.

O'Neill, G. and J. Dixon. 2005. "Nonresponse Bias in the American Time Use Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 10, 2005. 2958–2966. Minneapolis, MN: American Statistical

Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000193.pdf (accessed April 2018).

Roberts, C., N. Allum, and P. Sturgis. 2014. "Nonresponse and Measurement Error in an Online Panel. Does Additional Effort to Recruit Reluctant Respondents Result in Poorer Data Quality?" In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J. Krosnick, and P.J. Lavrakas, 337–362. Chichester: Wiley.

Rodgers, W.L., C. Brown, and G.J. Duncan. 1993. "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages." *Journal of the American Statistical Association* 88: 1208–1218. Doi: http://dx.doi.org/10.1080/01621459.1993.10476400.

Sakshaug, J.W. and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6: 113–122. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5094.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." *Public Opinion Quarterly* 74: 907–933. Doi: http://dx.doi.org/10.1093/poq/nfq057.

Sanders, D., H.D. Clarke, M.C. Stewart, and P. Whiteley. 2007. "Does Mode Matter for Modelling Political Choice? Evidence from the 2005 British Election Study." *Political Analysis* 15: 257–285. Doi: http://dx.doi.org/10.1093/pan/mpl010.

Sax, L.J., S.K. Gilmartin, and A.N. Bryant. 2003. "Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys." *Research in Higher Education* 44: 409–432. Doi: http://dx.doi.org/10.1023/A:1024232915870.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179: 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Statistisches Bundesamt. 2013. *Wirtschaftsrechnungen. Private Haushalte in der Informationsgesellschaft – Nutzung von Informations – und Kommunikationstechnologien*. Wiesbaden, Germany: Statistisches Bundesamt.

Stephenson, L.B. and J. Crête. 2011. "Studying Political Behavior: A Comparison of Internet and Telephone Surveys." *International Journal of Public Opinion Research* 23: 24–55. Doi: http://dx.doi.org/10.1093/ijpor/edq025.

Vannieuwenhuyze, J., G. Loosveldt, and G. Molenberghs. 2010. "A Method for Evaluating Mode Effects in Mixed-Mode Surveys." *Public Opinion Quarterly* 74: 1027–1045. Doi: http://dx.doi.org/10.1093/poq/nfq059.

Yeager, D.S., J.A. Krosnick, L. Chang, H.S. Javitz, M.S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75: 709–747. Doi: http://dx.doi.org/10.1093/poq/nfr020.

Received April 2016
Revised May 2018
Accepted May 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 117–135, http://dx.doi.org/10.2478/JOS-2019-0006

Cross-National Comparison of Equivalence and Measurement Quality of Response Scales in Denmark and Taiwan

Pei-shan Liao[1], *Willem E. Saris*[2], *and Diana Zavala-Rojas*[3]

The split-ballot multitrait-multimethod (SB-MTMM) approach has been used to evaluate the measurement quality of questions in survey research. It aims to reduce the response burden of the classic MTMM design, which requires repeating alternative formulations of a survey measure to the same respondent at least three times, by using combinations of two methods in multiple groups. The SB-MTMM approach has been applied to the European Social Survey (ESS) to examine the quality of questions across countries, including the differences in response design and measurement errors. Despite wide application of the SB-MTMM design in Europe, it is yet unknown whether the same quality of survey instruments can be achieved in both a different cultural context and in a logographic writing system, like the one in Taiwan.

This study tests for measurement invariance and compares measurement quality in Taiwan and Denmark, by estimating the reliability and validity of different response scales using the SB-MTMM approach. By using the same questions as in the ESS, a cross-cultural comparison is made, in order to understand whether the studied response scales perform equally well in Taiwan, compared to a European country. Results show that quality estimates are comparable across countries.

Key words: Split-ballot MTMM; reliability; validity; question quality.

1. Introduction

Survey measures take various forms, and studying their quality is an important issue, as they result in measurement error biases. For example, questions about subjective concepts can be measured by Likert-type response scales with different numbers and labels of response categories, with a feeling thermometer or using rating scores, among others (Alwin 1997; Schaeffer and Presser 2003). In comparative survey research, different measurement designs influence the response distributions and may lead to comparability problems across countries (see Bjørnskov 2010). Among the tools used to evaluate measurement quality, two approaches have been rather popular: the split-ballot experiment

[1] Center for Survey Research, RCHSS, Academia Sinica, 128 Academia Road, Sec. 2, Nangang Dist., Taipei 11529, Taiwan. Email: psliao@gate.sinica.edu.tw
[2] Sociometric Research Foundation, Carer Josep Pla 27 9-4, 08019 Barcelona, Spain. Email: w.saris@telefonica.net
[3] RECSM, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Edifici Mercè Rodoreda 24, 08005 Barcelona, Spain

Acknowledgments: This study was based on the Taiwan Social Change Survey, supported by the Ministry of Science and Technology, Taiwan (MOST 104-2420-H-001-005-SS3).

and the multitrait-multimethod (MTMM) approach (Saris et al. 2010; Saris et al. 2004). The basic principle of the split-ballot experiment approach is to randomly divide the respondents into two or more equal-sized subsamples with equal representativeness of the total sample (Schuman and Presser 1981; Petersen 2008). The respondents of each subsample answer survey questions simultaneously and under the same conditions. Variations in the questionnaire for each of the subsamples are treated as an experimental stimulus to examine questionnaire effects.

Alternatively, Campbell and Fiske (1959) suggested the MTMM design to evaluate the validity of social science concepts based on the correlations among measures of variables (Alwin 1974). The classic MTMM approach requires a respondent to answer questions about a minimum of three *traits,* that is, concepts or constructs measured using three different methods, for example, response scales, leading to nine different observed variables (Saris and Gallhofer 2014). Given the matrix, criteria for convergent and discriminant validity of these variables are advanced in Campbell and Fiske (1959) to assess validity. Structural equation modeling (SEM) can be applied to estimate the reliability and validity of each method. A comparison of fit statistics indicates which model best fits the data. Since the respondents need to repeatedly answer similar questions, it becomes a burden for them and may cause memory bias or order effect of the questions.

Saris et al. (2004) developed an approach to reduce the response burden by means of using different combinations of two methods in multiple groups. They combine the use of multiple groups in a split-ballot design, while the MTMM approach allows estimating the reliability and validity of the different questions. Such a split-ballot MTMM (SB-MTMM) approach has been applied to the European Social Survey (ESS) to examine the measurement quality of questions across countries, including the differences in response design and measurement errors (Oberski et al. 2007, 2010; Saris and Gallhofer 2014; Saris et al. 2008; Saris et al. 2010). Information about the quality of more than 2,700 questions from different European countries and the United States are stored in an online database in the Survey Quality Predictor (SQP) 2.1, which is an online system for survey quality prediction (Saris and Gallhofer 2014; Saris et al. 2011). On the basis of the data collected in all these countries using mainly English and European languages, a meta-analysis has been performed to develop a procedure to predict the quality of survey questions. This prediction tool is available in SQP 2.1. In the meta-analysis (Saris et al. 2011), it has been found that not only question characteristics, such as question wordings, response scales and labelling, but also the written and spoken language used in formulating the questions, determine the reliability and validity of questions.

With respect to nonWestern languages, some studies have evaluated different designs of response scales by the means of a split-ballot experiment (Lau 2016; Liao 2014). Some, such as Chen (2005) and Hsiao and Tu (2012), have applied MTMM to evaluate validity and reliability in Taiwan using Chinese-language content, but none had a focus on the effect of the formulation of single questions. Because no MTMM experiments have been done in Asia and specifically in Chinese, so far, the quality of survey questions cannot be predicted with SQP. Therefore, we designed this research to start by collecting quality estimates based on MTMM experiments in Taiwan. Previous studies have indicated that respondents in East Asian countries, for example, tend to more frequently choose responses in the middle of the scale than those in the West because of

the influence of collectivism (Chen et al. 1995; Harzing 2006). It is unknown whether the same quality of survey instruments can be achieved in both a different cultural context and writing system. The cultural transportability of experimental and pretesting techniques cannot be assumed; it has to be tested. For instance, Goerman and Caspar (2010) and Pan et al. (2005) have found that cognitive interviewing does not work equally well across cultures.

Using data from the SB-MTMM design in the Taiwan Social Change Survey (TSCS) and corresponding data from the 2002 ESS Round 1 in Denmark, this study aims to compare the measurement quality of different response designs across countries. It is of interest to explore the similarity, as well as differences, when the same experimental approach is applied. Therefore, we conduct a test for measurement invariance with the aim of concluding whether relationships and means across countries can be compared. The next section discusses the SB-MTMM approach that is used for this study and briefly introduces the test for measurement invariance. We then present the research design and results. Discussion on the findings are provided.

2. The SB-MTMM Design

A drawback of the classic MTMM design is the burden on each respondent of being asked multiple questions that assess the same construct. In addition, early questions may influence answers to later questions due to memory that is carried over. Consequently, the data quality may be overestimated (Saris et al. 2004). In order to minimize the carry-over effect from the previous answer, an interval of at least 20 minutes between the administration of the related items (Van Meurs and Saris 1990) is suggested.

The SB-MTMM design reduces the cognitive burden on respondents by using two, rather than three, methods in the MTMM design, while three traits are measured. Random samples of the same population are also used, as in the split-ballot experiments, but each respondent only needs to answer the questions concerning the same trait twice. This is seen to combine the benefits of the split-ballot approach and the MTMM approach, in that it enables researchers to evaluate measurement bias, reliability, and validity simultaneously, while reducing response burden (Saris and Gallhofer 2014).

We assume that the estimation model for the SB-MTMM design is the same as in the standard approach, given that the random samples are drawn from the same population. In the standard MTMM design, a minimum of three traits are measured using three different methods, leading to nine different observed variables. Therefore, a correlation matrix of 9×9 is obtained. However, this is not always the case when using the SB-MTMM approach. Nevertheless, the same models can be used, as we will show later. Various models have been suggested for analysis of the correlation matrices, and a true score model proposed by Saris and Andrews (1991) is commonly applied. The advantage of this model is that its standardization of the coefficients directly provides the estimates of the reliability and validity coefficients (Saris and Gallhofer 2014). Recent applications include those by Revilla and Saris (2013), Saris et al. (2010), Zavala-Rojas et al. (2018), Revilla (2015), Revilla et al. (2015), and Oberski et al. (2007).

The use of a minimum of three traits to be repeated using at least three methods serves the purpose of identification. With such a consideration, the model can be defined by the

following Equations (Saris and Andrews 1991):

$$Y_{ij} = r_{ij}T_{ij} + e_{ij} \tag{1}$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j \tag{2}$$

where Y_{ij} is the observed variable for the ith trait and the jth method; r_{ij} and v_{ij} are the reliability and validity coefficients for the ith trait and the jth method, respectively; T_{ij} is the true score or systematic component of the response Y_{ij}; e_{ij} is the random error associated with Y_{ij}; F_i is the ith trait (or factor); M_j is the variation in scores due to the jth method; and m_{ij} is the method effect for the ith trait and the jth method. The model posits that the observed variable is the sum of the systematic component plus a random error. Also, the systematic component of a response is the sum of the trait and the effect of the method used to assess it.

We make the assumption that the traits are correlated with one another. The random errors are not correlated with one another, nor with the independent variables in the different equations. The method factors are assumed to not be correlated with one another, nor with the traits or the random errors. Figure 1 is a graphical presentation of the true score model.

When all variables other than e_{ij} are standardized, v_{ij}, and m_{ij} correspond to the reliability, validity, and method effect coefficients, respectively, of a measure, while the squares of these coefficients present the reliability, validity and the method variance, respectively. In this approach, the reliability and validity of single questions have been evaluated, not complex concepts. As a result, the validity does not indicate how well the measured indicator represents the concept of interest. The validity is only affected by the method used, that is, $v_{ij}^2 = 1 - m_{ij}^2$. The lack of reliability will decrease the correlations between the variables, while the method effects will increase the correlations between the variables measured by the same or similar methods. This effect is called "common method variance". The model specified in Equation 1 and Equation 2 assumes that the disturbance term only contains a random error component, e_{ij}. Therefore, in this model, we make

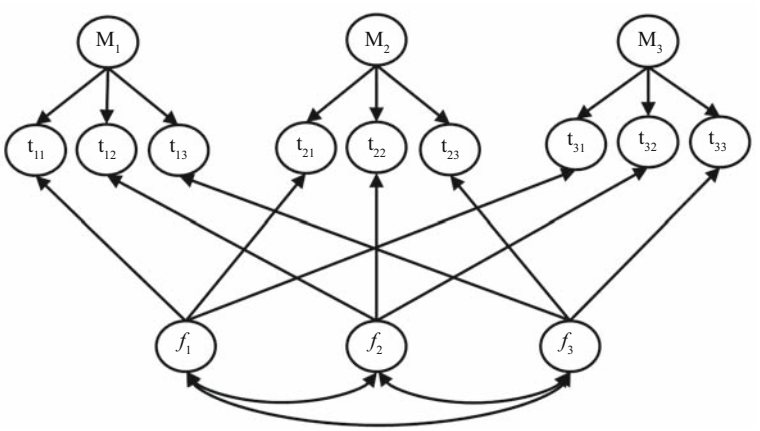

Fig. 1. MTMM model illustrating the true scores and their factor of interest.

the assumption that there is no unique component (Saris and Andrews 1991, 579). This assumption is plausible when the stem of the questions remains the same in the multitrait-multimethod experiment, and the variation only comes from the variations in the methods.

The total quality of a measure can then be computed as $q_{ij}^2 = r_{ij}^2 \times v_{ij}^2$, where q_{ij}^2 represents the amount of the variance of the observed variable, which is explained by the latent trait of interest. The quality indicators, reliability, and validity are typical measures of quality, which vary between zero and one, like correlation coefficients. With respect to the multiple groups in the SB-MTMM design, estimates for the parameters of the model can be obtained using structural equation modeling for multiple-group analysis (Saris and Gallhofer 2007, 2014).

3. Test for Measurement Invariance

With the SB-MTMM model, the variance-covariance matrix of the traits, F_i, is obtained. This correlation matrix is corrected for measurement error and can be used to test whether the same construct is measured across countries. The test for measurement invariance is typically done using the variance-covariance matrix of observed variables, although a criticism that has been referred to as *susceptibility,* that is, to what extent the procedure is sensitive to artifacts in the response process, is commonly made (Butts et al. 2006; Marsh and Byrne 1993; Byrne and Watkins 2003; Saris and Gallhofer 2014). Saris and Gallhofer (2014 chap. 16) showed that in a test for measurement invariance, the *response* process can be distinguished from the *cognitive* process. As we have said above, the variance-covariance matrix corrected for measurement errors will be obtained in the MTMM analysis, and this matrix can be used to test for the cognitive equivalence or comparability of the concepts in the different countries.

Therefore, we used the variance-covariance matrix of the latent traits to test for measurement invariance. The test is usually conducted in three steps, where each step is a prerequisite of the next one. In the first step, a *configural* model is fitted to check whether the pattern of fixed and free loadings and disturbance terms is the same across groups (Horn and McArdle 1992). In the second step, *metric invariance*, the configural model is restricted to one where the factor loadings of equivalent manifest variables are invariant across countries. When the model is not rejected, comparisons of relationships across groups can be made (Horn and McArdle 1992). The third step, *scalar invariance,* implies that, in addition to invariance in the factor loadings, intercepts of equivalent manifest variables are also restricted to be the same across groups. If the model is not rejected, comparisons of means can also be made across groups.

Figure 2 shows the path diagram of the model to test for measurement invariance. The model is specified in Equations (3) to (5).

$$f_1 = \tau_1 + \eta_1\lambda_1 + d_1 \tag{3}$$

$$f_2 = \tau_2 + \eta_1\lambda_2 + d_2 \tag{4}$$

$$f_3 = \tau_3 + \eta_1\lambda_3 + d_3 \tag{5}$$

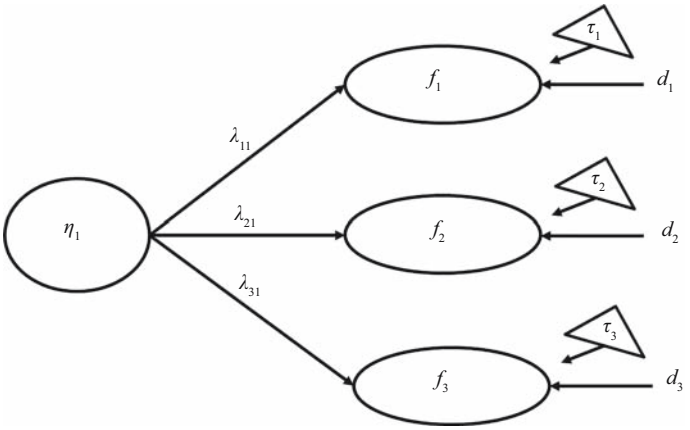

Fig. 2. Model to test for measurement invariance.

Where η_1 is the concept of interest and F1 to F3 represent the indicators used in the study corrected for measurement errors.

Standard restrictions were imposed to identify the model: the loading of the first trait (λ_1) and its corresponding intercept (d_1) were fixed to one and zero respectively. Secondly, we make the assumption that the error terms are not correlated with each other or with the latent variables. To test for metric invariance, we assume that the loadings (λ) are equal across groups, and for the scalar invariance we assume that the intercepts (d) are also equal across groups.

4. Research Design

Two data sources are used for this study, one from Taiwan and the other from Denmark, and both are collected using the computer-assisted personal interview (CAPI) technique. Taiwanese data are drawn from the 2015 Taiwan Social Change Survey (TSCS) (Fu et al. 2016), which included questions on globalization, work, family, and mental health, and included the SB-MTMM experiment. Surveys were delivered to randomly selected adults aged 18 years or older within each of the selected municipalities. A three-stage stratified sampling design was adopted based on the urbanization level and geographic areas of the townships and boroughs in Taiwan as the primary sampling unit (PSU). The probability proportional to size (PPS) sampling method was used in the first two stages – township and village or *li* under townships, respectively. Finally, household-registered residents in each village or *li*, which are equivalent-sized neighborhoods in urban areas, are systematically selected to obtain a representative sample of Taiwan's population. A total of 2,034 complete cases are obtained, with a response rate of 57%.

The experiment conducted in the 2015 TSCS adopted a two-group SB-MTMM design. The sample was randomly divided into two subsamples based on the respondent's number, which was assigned beforehand, as odd or even. The odd-numbered subsample (Sample 1) got Method 1 (M_1) first and then Method 3 (M_3), and the even-numbered subsample (Sample 2) got Method 2 (M_2) first, but Method 3 (M_3) next. As shown in Table 1, the combination of M_1 and M_2 was missing by design.

Table 1. Two-group SB-MTMM design.

	Method 1	Method 2	Method 3
Method 1	Sample 1		
Method 2	NONE	Sample 2	
Method 3	Sample 1	Sample 2	Sample 1 + 2

In other words, this set of correlations between the variables measured by M_1 and the variables measured by M_2 is absent. Saris et al. (2004) have shown, based on the work of Satorra (1993), that, in general, the parameters of this model evaluated with two groups are identified and all quality indicators can be estimated using multiple group estimation, except when the correlations between the traits are very similar or zero (Revilla and Saris 2013).

The measures for the SB-MTMM experiment include several questions. Each question is measured with two sets of response scales (M_1 and M_2 in the case of 2015 TSCS) that are answered by Sample 1 and Sample 2, respectively, and one other set (M_3) answered by all of the respondents. Both of the subsamples answer all other questions in the survey as well.

The questions used are commonly used indicators of the latent concept "Political satisfaction". The following three indicators of political satisfaction are used for the experimental design as follows:

1. On the whole, how satisfied are you with the present state of the economy in [country]?
2. Now thinking about the [country] government, how satisfied are you with the way it is doing its job?
3. And, on the whole, how satisfied are you with the way democracy works in [country]?

Using the same indicators for the three methods, M_1 is measured using a fully labeled four-point scale, with labels very satisfied, satisfied, dissatisfied and very dissatisfied. M_2 and M_3 are measured from 0 to 5 and from 0 to 10, respectively, both using show cards with only the endpoints labeled as "extremely dissatisfied" and "extremely satisfied". For all of the methods, a higher score indicates a higher level of satisfaction. The correlation matrices were obtained for analysis using a structural equation model. The design of the experiment has been summarized in Figure 3, where Ts_i denotes the Taiwan sample i, Ds_i denotes the Danish sample i where $i = 1, 2$, and c stands for the combination of the two samples within each country.

In order to estimate the parameters, covariance matrices obtained for the nine measures are used in the multi-group SEM in LISREL. The maximum likelihood (ML) approach is adopted to deal with missing data, which occurs by design (Saris et al. 2004).

The same measures of satisfaction, use of show cards and a two-group SB-MTMM experimental design can be found in the 2002 ESS Round 1. The data from Denmark (ESS1_DK) are used for the comparison with the data in Taiwan due to the same data collection mode of CAPI in both the main questionnaire and supplemental questions. The

Method / Question	M_1 (4-point scale) (01) Very dissatisfied (02) Fairly dissatisfied (03) Fairly satisfied (04) Very satisfied	M_2 (6-point scale) (00) Extremely dissatisfied ⋮ (05) Extremely satisfied	M_3 (11-point scale) (00) Extremely dissatisfied ⋮ (10) Extremely satisfied
Q1. How satisfied with present state of economy in country	TS1/DSC	TS2/DS1	TSC/DS2
Q2. How satisfied with the national government			
Q3. How satisfied with the way democracy works in country			

Fig. 3. Two-group SB-MTMM design for Denmark and Taiwan.

sampling design for ESS1_DK is a simple random sample based on a register-based sampling frame, with a lower age cut-off of 16 years. A total of 1,506 complete cases were obtained, with a response rate of 67.56%. More details can be found in the ESS1 – Documentation Report (European Social Survey 2014, 42–47). As in the case of Taiwan, the SB-MTMM experiment was performed alongside other questions, among others, about politics, work, family, well-being and immigration.

One difference in the experimental design between 2015 TSCS and ESS1_DK is that all the respondents in the latter got M_1 first, and then M_2 and M_3 for samples 1 and 2, respectively. Therefore, the combination of M_2 and M_3 is missing by design in the Danish data. The differences in the data structures are clearly observable in the correlation matrices presented in Table 3 and Table 4. Nevertheless, the same model can be estimated on the basis of these two different correlation matrices.

Another difference is that in Taiwan, an unfolding technique, in which interviewers first asked about the direction and then about the degree of attitudes (Schaeffer and Presser 2003), was used for M_1, with the scale coded reversely as 1 = very satisfied to 4 = very dissatisfied. In the ESS, one direct question was used, in which all four categories were presented immediately. Although the data have been recoded to have the same response order as that in the Danish data, the difference in procedure means that we were not able to determine the effect of the scale length, only because this effect is confounded with other aspects. However, it is possible to determine which measure is better in each country and across countries.

5. Results of the SB-MTMM Experiment

Socio-demographic variables were first compared between Denmark and Taiwan with post-stratified weights. As shown in Table 2, the distributions of demographic characteristics are similar in age and gender. The proportions of those aged 60 years or older are lower when compared to other age groups. Also, there are similar proportions of men and women in both samples. On the other hand, the proportions of married and widowed respondents in Taiwan are higher, but those of single or divorced respondents are higher in Denmark. It is noted that ESS used other variables to ask respondents whether they live with a partner, but a category of "cohabitant" is included in TSCS for marital

Table 2. Description of Denmark and Taiwan samples.

Country	Denmark		Taiwan	
Variable	*f*/M	%/SD	*f*/M	%/SD
Age				
16–29 years	304	20.2%	401	19.8%
30–39 years	277	18.4%	359	17.7%
40–49 years	245	16.3%	382	18.8%
50–59 years	294	19.5%	378	18.6%
60–79 years	197	13.1%	263	13.0%
70 years or older	189	12.5%	247	12.2%
P = .553	N = 1506	100%	N = 2030	100%
Gender				
Female	736	49%	1043	51.3%
Male	766	51%	991	48.7%
P = .096	N = 1502	100%	N = 2034	100%
Marital***				
Single, never married	486	31.3%	594	29.3%
Married	800	53.6%	1141	56.2%
Divorced	119	8.0%	123	6.1%
Separate	15	1.0%	9	0.4%
Widowed	91	6.1%	149	7.3%
Cohabitant	0	0%	14	0.7%
P = .000	N = 1493	100%	N = 2030	100%
Educational level***				
Elementary or less	26	1.7%	416	20.5%
Junior high schoold	392	26.2%	256	12.6%
Senior high school	733	49%	541	26.7%
Tertiary education or higher	344	23%	815	40.2%
P = .000	N = 1495	100%	N = 2028	100%
Health status***				
Very good	648	43.2%	468	23.0%
Good	512	34.2%	619	30.5%
Fair	253	16.9%	630	31.0%
Bad	86	5.7%	314	15.5%
P = .000	N = 1499	100%	N = 2031	100%
Interested in Politics***				
Very interested	202	13.4%	34	1.7%
Quite interested	723	48.1%	330	16.3%
Hardly interested	487	32.4%	654	32.4%
Not at all interested	90	6.0%	1003	49.6%
P = .000	n = 1502	100%	n = 2021	100%

Table 2. Continued.

Country	Denmark		Taiwan	
Variable	*f*/M	%/SD	*f*/M	%/SD
Method 1 (4-point scale)				
Q1. How satisfied with present state of economy in country***	2.91	0.577	2.12	0.771
Q2. How satisfied with the national government***	2.72	0.656	2.04	0.767
Q3. How satisfied with the way democracy works in country***	3.10	0.595	2.58	0.752
Method 2 (6-point scale)				
Q1. How satisfied with present state of economy in country***	3.46	0.917	1.96	1.241
Q2. How satisfied with the national government***	3.06	1.091	1.82	1.254
Q3. How satisfied with the way democracy works in country***	3.74	0.854	2.77	1.339
Method 3 (11-point scale)				
Q1. How satisfied with present state of economy in country***	6.92	1.938	3.94	2.036
Q2. How satisfied with the national government***	5.86	2.268	3.58	2.089
Q3. How satisfied with the way democracy works in country***	7.24	1.876	5.31	2.251

*p < .05.
**p < .01.
***p < .001.

status. If "cohabitant" is dropped, the proportions of other categories in marital status increase slightly, from 0.06% to 0.4%, in TSCS and the result of the Chi-square test remains significant. Educational levels are recoded for both ESS1_DK and TSCS for comparison. Almost half of the Danish sample have a senior high school degree (49%), while 40% of the respondents in the Taiwanese data have a higher tertiary education degree, including formal education at colleges, universities and higher degrees.

There are other variables that can be used to reveal the difference between Denmark and Taiwan, such as health status and interest in politics. Both ESS1_DK and 2015 TSCS employed five response categories for health status, but the former used a balanced scale, from "very good" to "very bad" with "fair" as the middle response, while the latter used an unbalanced one, from "excellent" to "bad." The categories of "bad" and "very bad" in ESS1_DK are combined and so are "excellent" and "very good" in 2015 TSCS, resulting in four response categories (see Table 2).

With regard to interest in politics, both samples employed the same response categories for measurement. It is noticeable that more than 60% of the Danish sample indicated certain levels of interest in politics, while nearly half of the Taiwanese sample were not at all interested in politics.

Table 3. Correlations, means, and standard deviations of Danish samples[1].

Sample 1	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.410	1							
Q3	.288	.288	1						
M_2									
Q1	.414	.267	.179	1					
Q2	.262	.677	.162	.410	1				
Q3	.269	.261	.473	.407	.388	1			
M_3									
Q1	.0	.0	.0	.0	.0	.0	1		
Q2	.0	.0	.0	.0	.0	.0	.0	1	
Q3	.0	.0	.0	.0	.0	.0	.0	.0	1
Mean	2.91	2.73	3.11	6.94	5.92	7.31	.0	.0	.0
S.D.	.579	.652	.599	1.915	2.273	1.884	1.0	1.0	1.0
n	653	653	653	653	653	653			
Sample 2	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.497	1							
Q3	.411	.317	1						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.0	1				
Q3	.0	.0	.0	.0	.0	1			
M_3									
Q1	.554	.373	.300	.0	.0	.0	1		
Q2	.388	.744	.196	.0	.0	.0	.466	1	
Q3	.362	.309	.603	.0	.0	.0	.421	.372	1
Mean	3.46	3.08	3.74	.0	.0	.0	6.92	5.95	7.30
S.D.	.911	1.069	.844	1.0	1.0	1.0	1.962	2.203	1.794
n	687	687	687				687	687	687

[1]All of the correlation coefficients are significant at the .000 level.

Table 4. Correlations, means, and standard deviations of Taiwanese samples[1].

Sample 1	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.678	1							
Q3	.317	.388	1						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.0	1				
Q3	.0	.0	.0	.0	.0	1			
M_3									
Q1	.525	.518	.304	.0	.0	.0	1		
Q2	.496	.628	.360	.0	.0	.0	.733	1	
Q3	.252	.314	.532	.0	.0	.0	.428	.482	1
Mean	2.09	2.02	2.57	.0	.0	.0	4.07	3.68	5.41
S.D.	.749	.755	.757	1.0	1.0	1.0	1.963	2.039	2.219
n	880	880	880				880	880	880
Sample 2	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	.0	.0	.0						
Q2	.0	.0	.0						
Q3	.0	.0	.0						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.686	1				
Q3	.0	.0	.0	.381	.475	1			
M_3									
Q1	.0	.0	.0	.673	.637	.345	1		
Q2	.0	.0	.0	.589	.748	.399	.781	1	
Q3	.0	.0	.0	.293	.337	.651	.392	.419	1
Mean	.0	.0	.0	1.97	1.80	2.77	3.77	3.44	5.21
S.D.	1.0	1.0	1.0	1.216	1.239	1.331	2.067	2.100	2.298
n				890	890	890	890	890	890

[1]All of the correlation coefficients are significant at the .000 level.

As for the satisfaction measured by three response scales, significant differences are found between countries, as well as among methods. Among different methods, the mean scores of three satisfaction questions are higher in Denmark than in Taiwan. In particular, the differences between Denmark and Taiwan are larger when satisfaction is measured by M_2 and M_3, despite the consistently low levels of satisfaction in the Taiwanese sample.

The results of the satisfaction measures using a two-group SB-MTMM design are reported in Table 3 and Table 4 for ESS1_DK and 2015 TSCS, respectively, indicating incomplete data in each of the subsamples. Since both of the datasets employed approximately the same SB-MTMM experimental design, the parameter estimation followed the same procedure. The correlations for the unobserved variables are indicated by zeros and the variances by ones, as required for the multiple-group analysis with incomplete data in LISREL (Allison 1987). The correlation between the variables measured by M_1 and M_2 is missing by design. Therefore, the parameters are estimated based on the incomplete covariance matrix. In addition, in order to estimate the coefficients of reliability, validity, and method effects for the two randomly selected subsamples simultaneously, we make the assumption that the model is the same for both groups, except for the specification of selecting the variables of the two groups.

The Taiwanese data had the peculiarity that the correlations between questions 1 and 2 were much higher than the correlation between these variables and question 3. The program Jrule (Van der Veld et al. 2008), which was used to detect misspecifications in the model, also detected this high correlation and suggested introduction of a correlated error between questions 1 and 2 in the model for the Taiwanese data. Only in Method 2 was this correlated error not significantly different from zero. In the other two methods, these correlations were 0.14 for M_1 and 0.28 for M_3. The explanation is not so simple, but it is clear that a deteriorating economy has been a serious issue in Taiwan in the past decade, and this has been seen as the responsibility of the government. Research in political science has indicated such consequences of economic performance on voting behavior and named it "economic voting" (Wu and Lin 2013). One can therefore expect a much higher correlation between satisfaction with the government and satisfaction with the economy than between these two and the functioning of democracy, which does not depend so much on the present government only. With this one correction, a proper solution is obtained with a $chi^2 = 32.20$ and $df = 38$ after we corrected for the zero cells in the correlation matrices, the RMR = .011. The Jrule approach to test for local misspecifications (Saris et al. 2009) did not suggest improvements.

Table 5. Estimates of the parameters for the two-group SB-MTMM design[1].

	Reliability						Validity					
Method	M_1		M_2		M_3		M_1		M_2		M_3	
Country	D	T	D	T	D	T	D	T	D	T	D	T
Q1	.55	.55	.74	.72	.58	.72	.79	.86	.92	.98	.74	.81
Q2	.74	.69	.90	.87	.77	.79	.87	.90	.94	.98	.81	.83
Q3	.91	.62	.62	.85	.76	.62	.76	.88	.90	.98	.81	.81
Average	.73	.62	.75	.81	.70	.71	.81	.88	.92	.98	.79	.82

[1]"D" denotes the ESS1_DK data and "T" denotes the 2015 TSCS.

Table 6. The quality of the different questions for the different methods used[1].

	M_1		M_2		M_3	
Country	D	T	D	T	D	T
Q1	0.43	0.47	0.68	0.71	0.45	0.58
Q2	0.64	0.62	0.85	0.85	0.62	0.66
Q3	0.69	0.54	0.55	0.83	0.62	0.50
Average	0.59	0.54	0.69	0.8	0.56	0.58

[1]"D" denotes the ESS1_DK data and "T" denotes the 2015 TSCS.

The estimated reliabilities and validities for ESS1_DK and 2015 TSCS are reported in Table 5. As the total quality of a measure is defined as follows:

$q_{ij}^2 = r_{ij}^2 \times v_{ij}^2$, Table 6 has been derived from Table 5.

Table 6 shows that in both countries, the second method, the six-point scale has better quality on average over the three questions. This is also true for three questions, except for question 3, in Denmark. The quality of the measures using Method 1 (four-point scale) and Method 3 (eleven-point scale) do not differ very much. However, when we look at the estimated validities of the questions using Method 1 and Method 3 we see that they are slightly better in Taiwan, while the reliability using Method 1 in Taiwan is lower.

5.1. *Results of the Test for Measurement Invariance*

Table 7 shows the variance-covariance matrix of the latent traits. As the baseline model has only recently been identified, it is not possible to conduct a robustness test with the configural model. When the loadings are restricted, Jrule shows that the loading of the second trait is misspecified. As was mentioned above, the government is seen as responsible for the economic situation. This seems to be stronger for the case in Taiwan

Table 7. The variance-covariance matrix of the traits.

	Denmark		
n = 1502	Q_1	Q_2	Q_3
Q_1	.17		
Q_2	.13	.34	
Q_3	.09	.09	.19
Mean	2.91	2.73	3.11
	Taiwan		
n = 2020	Q_1	Q_2	Q_3
Q_1	.28		
Q_2	.28	.37	
Q_3	.15	.19	.29
Mean	2.09	2.02	2.57

Table 8. Likelihood ratio test of the metric and scalar models.

	Chi-square	Chi-square difference	DF difference	Pr. ($>\text{Chi}^2$)
Partially metric invariant model	0.0766			
Partially scalar invariant model	0.8217	0.74511	1	0.388

than in Denmark. This means that this measurement instrument has only partial metric invariance, that is, only the first and the third items are comparable across countries. Leaving this loading free, the concepts can be seen as comparable across the two countries.

As equality of loadings is a prerequisite for scalar invariance, the intercepts are restricted to being equal in both countries, except in the second trait. The likelihood ratio test (Table 8) indicates that the fit of the scalar model is not significantly different from the one of the metric model, and Jrule did not show additional misspecifications. These results imply that, at the cognitive level, partial scalar invariance is established. As the observed data are corrected for measurement errors, this result means that the relationships between these concepts and other variables and the latent means of the concepts can be compared across countries.

6. Conclusion

The SB-MTMM approach has been widely applied in the ESS to evaluate the quality of survey measures. It remains unclear whether this approach performs equally well in logographic writing systems. In addition, it is of interest to explore possible similarity or difference between ESS and Taiwanese data. Using a two-group experimental design, Danish data from ESS Round 1 and the 2015 TSCS are compared. The results indicated that questions measured by six-point scales with labels at endpoints (M_2) have the best quality, while the measures on either a four-point scale with full labels or an eleven-point scale are equally acceptable. Although differences between Danish and Taiwanese data can be observed, the findings are comparable, despite the fact that the order of applying methods differed. These findings are contrary to previous research suggesting that fully labeled response scales provide higher reliability than those with endpoints (Alwin and Krosnick 1991; Holbrook et al. 2006; Weng 2004), while the results are consistent with other studies (Saris and Gallhofer 2014; Saris et al. 2004). The designs of response scale deserve further examination. However, these methods differed in more than just this one aspect, which may explain these results.

One possible reason for the relatively poor quality of M_1 in the 2015 TSCS may be the different measurement procedures used for the different methods during the face-to-face interview. An unfolding technique, in which interviewers first asked about the direction and then about the degree of attitudes (Schaeffer and Presser 2003), is used for M_1 to minimize the tendency of choosing the middle category. On the other hand, show cards are provided upon request for M_2 and M_3, so it is easier for the respondents to answer the questions using Methods 2 and 3, rather than using Method 1. While the inquiry process

should be considered as part of the methods, researchers need to be cautious with its influence on data quality.

One cannot draw general conclusions about the effect of different aspects of the methods on the quality of questions, because often, like here, more aspects vary at the same time. Also, findings on quality of measures may differ by the measured topics. For general conclusions, we refer to the results of meta-analyses over large numbers of MTMM experiments (Saris and Gallhofer 2014).

A second result is that the concept "political satisfaction" is only partially invariant across the two countries. The results of the invariance test show that the understanding of the indicators of satisfaction with the economy and with the way democracy works are comparable in Denmark and Taiwan. However, this is not the case for satisfaction with the government. For this last indicator, there seems to be a different interpretation in the two countries. This signifies that means and relationships of the latent variable "political satisfaction" can be compared across countries, but composite scores can only be compared if one uses only the comparable indicators in computing the composite scores.

7. References

Allison, P.D. 1987. "Estimation of Linear Models with Incomplete Data." In *Sociological Methodology*, edited by C.C. Clogg, 71–103. Washington DC: American Sociological Association. Doi: https://doi.org/10.2307/271029.

Alwin, D.F. 1974. "Approaches to the Interpretation of Relationships in the Multitrait Multimethod Matrix." *Sociological Methodology* 5: 79–105. Doi: https://doi.org/10.2307/270833.

Alwin, D.F. 1997. "Feeling Thermometers versus 7-point Scales: Which Are Better?" *Sociological Methods and Research* 25: 318–340. Doi: https://doi.org/10.1177/0049124197025003003.

Alwin, D.F. and J.A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20: 139–181. Doi: https://doi.org/10.1177/0049124191020001005.

Bjørnskov, C. 2010. "How Comparable Are The Gallup World Poll Life Satisfaction Data?" *Journal of Happiness Studies* 11: 41–60. Doi: https://doi.org/10.1007/s10902-008-9121-6.

Butts, M.M., R.J. Vandenberg, and L.J. Williams. 2006. "Investigating the Susceptibility of Measurement Invariance Tests: The Effects of Common Method Variance." *Academy of Management Proceedings* 2006(1): D1–D6. Doi: https://doi.org/10.5465/AMBPp.2006.27182126.

Byrne, B.M. and D. Watkins. 2003. "The Issue of Measurement Invariance Revisited." *Journal of Cross-Cultural Psychology* 34(2): 155–175. Doi: https://doi.org/10.1177/0022022102250225.

Campbell, D.T. and D.W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2): 81–105. Doi: https://doi.org/10.1037/h0046016.

Chen, C.K. 2005. "Construct Model of Knowledge: Based Economy Indicators." *Management Review* 24(3): 17–41. Doi: https://doi.org/10.6656/MR.2005.24.3.CHI.17.

Chen, C., S.Y. Lee, and H.W. Stevenson. 1995. "Response Style and Cross-Cultural Comparisons of Rating Scales among East Asian and North American Students." *Psychological Science* 6: 170–175. Doi: https://doi.org/10.1111/j.1467-9280.1995.tb00327.x.

ESS Round 1: European Social Survey. 2014. *ESS-1 2002 Documentation Report*. Edition 6.4. Bergen, European Social Survey Data Archive, NSD – Norwegian Centre for Research Data for ESS ERIC. Available at: http://www.europeansocialsurvey.org/docs/round1/survey/ESS1_data_documentation_report_e06_4.pdf (accessed May 2016).

Fu, Y.-C., Y.-H. Chang, S.-H. Tu, and P.-S. Liao. 2016. *2015 Taiwan Social Change Survey (Round 7, Year 1): Globalization, Work, Family, Mental Health, and Political Participation* (C00315_2) [Data file]. Available at Survey Research Data Archive, Academia Sinica. Doi: https://doi.org/10.6141/TW-SRDA-C00315_1-1.

Goerman, P.L. and R.A. Caspar. 2010. "Managing the Cognitive Pretesting of Multilingual Survey Instruments: A Case Study of Pretesting of the U.S. Census Bureau Bilingual Spanish/English Questionnaire." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. Harkness, et al.: 75–90. John Wiley and Sons, Inc. Doi: https://doi.org/10.1002/9780470609927.ch5.

Harzing, A.W. 2006. "Response Styles in Cross-National Survey Research: A 26 Country Study." *International Journal of Cross Cultural Management* 6 (2)(August 1): 243–266. Doi: https://doi.org/10.1177/1470595806066332.

Hsiao, C.-C. and C.-H. Tu. 2012. "Common Method Variance in the Measurement of Teachers' Creative Teaching." *Psychological Testing* 59(4): 609–639. Doi: http://dx.doi.org/10.7108%2fPT.201212.0609.

Holbrook, A., Y.K. Cho, and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70: 565–595. Doi: https://doi.org/10.1093/poq/nfl027.

Horn, J.L. and J.J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18(3–4): 117–144. Doi: https://doi.org/10.1080/03610739208253916.

Lau, C.Q. 2016. "Rating Scale Design among Ethiopian Entrepreneurs: A Split-Ballot Experiment." *International Journal of Public Opinion Research* edw031. Doi: https://doi.org/10.1093/ijpor/edw031.

Liao, P.-S. 2014. "More Happy or Less Unhappy? Comparison of the Balanced and Unbalanced Designs for the Response Scale of General Happiness." *Journal of Happiness Studies* 15(6): 1407–1423. Doi: https://doi.org/10.1007/s10902-013-9484-1.

Marsh, H.W. and B.M. Byrne. 1993. "Confirmatory Factor Analysis of Multitrait-Multimethod Self-concept Data: Between-group and Within-group Invariance Constraints." *Multivariate Behavior Research* 28(3): 313–449. Doi: https://doi.org/10.1207/s15327906mbr2803_2.

Van Meurs, A. and W.E. Saris. 1990. "Memory Effects in MTMM Studies." In *Evaluations of Measurement Instruments by Metaanalysis of Multitrait-Multimethod Studies*, edited by W.E. Saris and A. van Meurs, 134–146. Amsterdam: North Holland.

Oberski, D.L., W.E. Saris, and J. Hagenaars. 2007. "Why Are There Differences in Measurement Quality across Countries?" In *Measuring Meaningful Data in Social Research*, edited by G. Loosveldt and Swyngedouw. Leuven: Acco. Available at:

http://daob.nl/wp-content/uploads/2013/03/Oberski-Saris-Why-are-there-differences-in-measurement-quality-across-countries.pdf (accessed January 2019).

Oberski, D., W.E. Saris, and J.A. Hagenaars. 2010. "Categorization Errors and Differences in the Quality of Questions in Comparative Surveys." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. Harkness, et al.: 435–453. Hoboken, NJ: Wiley. Doi: https://doi.org/10.1002/9780470609927.ch23.

Pan, Y., B. Craig, and S. Scollon. 2005. "Results from Chinese Cognitive Interviews on the Census 2000 Long Form: Language, Literacy, and Cultural Issues." *Statistical Research Division's Research Report Series* (Survey Methodology 2005 – 09). Washington, DC: U.S. Bureau of the Census. Available at https://www.census.gov/srd/papers/pdf/rsm2005-09.pdf (accessed November 2017).

Petersen, T. 2008. "Spilt Ballot as An Experimental Approach to Public Opinion Research." In *The Sage Handbook of Public Opinion Research*, edited by W. Donsbach and M.W. Traugott, 322–329. Los Angeles, CA: Sage. Available at: http://methods.sagepub.com/book/sage-hdbk-public-opinion-research/n30.xml (accessed January 2019).

Revilla, M. 2015. "Comparison of the Quality Estimates in a Mixed-Mode and a Unimode Design: An Experiment from the European Social Survey." *Quality and Quantity* 49(3): 1219–1238. Doi: https://doi.org/10.1007/s11135-014-0044-5.

Revilla, M. and W.E. Saris. 2013. "The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems." *Structural Equation Modeling: A Multidisciplinary Journal* 20: 27–46. Doi: https://doi.org/10.1080/10705511.2013.742379.

Revilla, M., W.E. Saris, G. Loewe, and C. Ochoa. 2015. "Can a Non-Probabilistic Online Panel Get Similar Question Quality as the ESS?" *International Journal of Market Research* 57(3): 395–412. Available at: https://www.mrs.org.uk/ijmr_article/article/104501 (accessed January 2019).

Saris, W.E. and F.M. Andrews. 1991. "Evaluation of Measurement Instruments Using a Structural Modeling Approach." In *Measurement Errors in Surveys*, edited by P.P. Biemer, et al.: 575–597. New York, NY: Wiley.

Saris, W.E. and I.N. Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 1: 29–43. Doi: http://dx.doi.org/10.18148/srm/2007.v1i1.49.

Saris, W.E. and I.N. Gallhofer. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (Second edition). Hoboken, NJ: Wiley.

Saris, W.E., A. Satorra, and G. Coenders. 2004. "A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design." *Sociological Methodology* 34: 311–347. Doi: https://doi.org/10.1111/j.0081-1750.2004.00155.x.

Saris, W.E., A. Satorra, and W.M. van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling: A Multidisciplinary Journal* 16(4): 561–582. Doi: https://doi.org/10.1080/10705510903203433.

Saris, W.E., R. Veenhoven, A.C. Scherpenzeel, and B. Brunting. 2008. *A Comparative Study of Satisfaction with Life in Europe*. Budapest: Eötvös University Press.

Saris, W.E., M. Revilla, J.A. Krosnick, and E.M. Shaffer. 2010. "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response

Options." *Survey Research Methods* 4: 61–79. Doi: https://doi.org/10.18148/srm/2010.v4i1.2682.

Saris, W., D. Oberski, M. Revilla, D. Zavala, L. Lilleoja, I. Gallhofer, and T. Gruner. 2011. "The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions." RECSM Working Paper 24, Universitat Pompeu Fabra. Available at: https://www.upf.edu/documents/3966940/3986764/RECSM_wp024.pdf (accessed January 2019).

Satorra, A. 1993. "Asymptotic Robust Inferences in Multi-sample Analysis of Augmented Moment Matrices." In *Multivariate Analysis; Future Directions*, edited by R. Rao and C.M. Cuadras, 211–229. Amsterdam, North Holland.

Schaeffer, N.C. and S. Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29: 65–88. Doi: https://doi.org/10.1146/annurev.soc.29.110702.110112.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Van der Veld, W., W.E. Saris, and A. Satorra. 2008. *Jrule 2.0: User Manual*, Unpublished document.

Weng, L-J. 2004. "Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability." *Educational and Psychological Measurement* 64: 956–972. Doi: https://doi.org/10.1177/0013164404268674.

Wu, C-E. and Y-T. Lin. 2013. "Cross-Strait Economic Openness, Identity, and Vote Choice: An Analysis of the 2008 and 2012 Presidential Elections." *Journal of Electoral Studies* 20(2): 1–36. Doi: https://doi.org/10.6612/tjes.2013.20.02.01-35.

Zavala-Rojas, D., R. Tormos, W. Weber, and M. Revilla. 2018. "Designing Response Scales with Multi-Trait-Multi-Method Experiments." *Mathematical Population Studies* 25(2): 66–81. Doi: https://doi.org/10.1080/08898480.2018.1439241.

Received October 2016
Revised July 2018
Accepted July 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 137–165, http://dx.doi.org/10.2478/JOS-2019-0007

An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics

Jack Lothian[1], *Anders Holmberg*[2], *and Allyson Seyb*[3]

The linking of disparate data sets across time, space and sources is probably the foremost current issue facing Central Statistical Agencies (CSA). If one reviews the current literature looking for the prevalent challenges facing CSAs, three issues stand out: 1) using administrative data effectively; 2) big data and what it means for CSAs; and 3) integrating disparate data set (such as health, education and wealth) to provide measurable facts that can guide policy makers. CSAs are being challenged to explore the same kind of challenges faced by Google, Facebook, and Yahoo, which are using graphical/semantic web models for organizing, searching and analysing data. Additionally, time and space (geography) are becoming more important dimensions (domains) for CSAs as they start to explore new data sources and ways to integrate those to study relationships. Central agency methodologists are being pushed to include these new perspectives into their standard theories, practises and policies. Like most methodologists, the authors see surveys and the publications of their results as a process where estimation is the key tool to achieve the final goal of an accurate statistical output. Randomness and sampling exists to support this goal, and early on it was clear to us that the incoming "it-is-what-it-is" data sources were not randomly selected. These sources were obviously biased and thus would produce biased estimates. So, we set out to design a strategy to deal with this issue.

This article presents a schema for integrating and linking traditional and non-traditional datasets. Like all survey methodologies, this schema addresses the fundamental issues of representativeness, estimation and total survey error measurement.

Key words: Representativeness; timeline databases; statistical registers; Estimation; administrative data.

1. Introduction

The linking of disparate data sets across time, space and sources is probably the foremost current issue facing Central Statistical Agencies CSAs. If one reviews the current literature looking for the prevalent challenges facing CSAs, three issues stand out: 1) using administrative data effectively; 2) big data and what it means for CSAs; and 3) integrating disparate data sets (such as health, education and wealth) to provide measurable facts that can guide policymakers. CSAs are being challenged to explore the same kind of concerns facing Google and Facebook, which are using graphical/semantic web models Ferrara et al. 2011 for organizing, searching and analyzing data. Additionally, time and space

[1] 360 Hinton Ave S, Ottawa ON K1Y1A5 Canada. Email: lothianjack@netscape.net
[2] Statistics Norway, Division for Methodology, Akersveien 26 Oslo, Norway. Email: anders.holmberg@ssb.no
[3] Stats NZ, Statistical Methods, Private Bag 4741, Christchurch 8011, New Zealand. Email: allyson.seyb@stats.govt.nz

(geography) are becoming more important dimensions (domains) for CSAs as they start to explore causal models. Central agency methodologists are being pushed to include these new perspectives into their standard theories, practises and policies. This article presents a schema for integrating and linking traditional and nontraditional data sets. Like all survey methodologies, this schema addresses the fundamental issues of representativeness, estimation and total survey error measurement.

Over the past decade, a new design paradigm has emerged concerning strategies for integrating disparate data sets to provide new understandings from the data. The development of this paradigm is currently not focused and there are multiple paths of advancement being pursued, such as big data, evolutionary databases (Fowler and Sadalage 2003), semantic web models, graphical query databases and many others. In a Graphical Queries Database (GQD), every element has a direct pointer to its adjacent elements. The simplest GQD pointer is a first order tree of one-to-one links. Semantic web links are first-order trees where the linkage function (the verb) becomes a generalized function. All these areas of research have a core issue: combining/linking large disparate data sets in a feasible and cost-effective manner. The complexity of the information, the fuzziness of the data inside each data set, the fuzziness of the linkage strategies, the large number of disparate data set, covering disjoint populations, the lack of control of the content and quality of administrative data, and the size of the data sets preclude the use of many straightforward classical solutions (Baker et al. 2013; Bakker and Daas 2012; Hand 2018; Holt 2000; Zhang 2012).

All the above-mentioned strategies appear to follow parallel paths to the same general solution. All these approaches propose viewing disparate data set integration as an evolutionary or ongoing process. The data and the database structure evolve as new information is added; as more data sets are added and linked; as new relationships between data sets are discovered and added; as new models of how different data sets interact are discovered; as new editing rules and methodologies are found; as the questions that we want answered change; as we become more knowledgeable of the data; and so on. The evolutionary nature of the problem implies that no fixed solution can succeed over an extended period. All the strategies cited above embrace evolution and make it part of the solution.

The core data design concept we are proposing is a simplified adaption of how many online search companies structure and search data. The major point of departure of our schema versus these online solutions is our inclusion of time. For these companies, the point in time at which the measurements are made is not usually a relevant characteristic, but for our schema, it will be a fundamental aspect of the data. Later in the document, we will also see that "space" or geography will become a necessary dimension of our design schema. Our schema will be underpinned and anchored by a space-time lattice, through which our entities will travel. It is somewhat akin to the game called "Life". We will call the structured collection of common files (administrative, survey, register or census) an evolutionary schema. The term "evolutionary" implies that the database constructs entities' event timelines and these timelines are updated with new current events. The event timelines evolve. In the paradigm of database design and programming, "evolutionary database" design has a different sense. It is the database design schema and algorithms that are always evolving in an incremental fashion. Our proposed design will be evolutionary in this sense as well.

We present a conceptual schema for dealing with the integration of nontraditional and traditional survey data sets. It is important to note that we will be presenting a strategy for structuring, analyzing problems and answering questions, rather than a specific solution. As in classical survey design, our final goal will be a strategy to provide the best possible estimates. To achieve this, we convey the message that methodologists must understand the whole process that will produce the estimates, not just focus on one phase of the process.

We believe that the basis for understanding this process and creating interpretable and meaningful estimates will be a system of statistical base registers, plus consistent monitoring and maintenance strategies. These statistical registers serve as lighthouses for illuminating 'trusted' estimation procedures and provide a benchmark for comparing and investigating representativeness concerns. We believe that our schema provides a broad and general framework for CSAs working with large collections of administrative data and other conveniently available data sets/databases that we refer to as "it-is-what-it-is" data sets. We offer a framework for: structuring the non-probabilistic data; making it useful for cause and effect statistical inference; incrementally developing, designing and maintaining the database system; and, inserting total survey error concepts into the schema. Our schema does not provide detailed designs for these processes, instead we provide a pseudo-scientific framework for addressing survey design questions when using non-probabilistic data sets.

Section 2 presents the concept of "it-is-what-it-is" data sources. Section 3 discusses the importance of registers to support estimation and eliminate potential biases within the database schema. Section 4 presents an overview of our data model for structuring and using the data in the evolutionary database. Section 5 discusses how estimation might take place in the evolutionary schema. Section 6 discusses the place of metadata, and measuring Total Survey Error (TSE) and controlling quality in this evolutionary schema. Section 7 is a summary.

2. Structured Framework for Using "it-what-it-is" Data Sets

2.1. "It-is-what-it-is" Data Sets

In our evolutionary schema, we will assume that all data sources integrated into the evolutionary database are provided by an outside agency that is beyond the control and influence of the owners of the evolutionary database. These outside sources could be administrative files, censuses, registers, client lists, commercial transactions, sensor readings, survey files, sample files, and so on. Our source data sets will be what Sharon Lohr (Lohr et al. 2015) recently referred to as "it-is-what-it-is" data sets. "It-is-what-it-is" data sets are source files where the survey methodologist has no control over the selection probabilities, nor the content of the files. It should be noted that the true sample selection probabilities for the entities in these external sourced data sets may be non-probabilistic and/or unknowable.

As expressed by Sharon Lohr, the term "it-is-what-it-is" has a wider sense. As survey methodologists, we may be asked to answer questions where the sole source of information concerning these questions are "it- is-what-it-is" data sets and thus we may be

forced to use these data sets despite their limitations. In this case, methodologists must resort to pseudo-scientific methods to address the questions. If this is the case, Sharon says methodologists need to be aware of the data sets' limitations or "what it is". Sharon stated that "it-is-what-it-is" data sets fundamentally change our analysis paradigm and we need to understand this point. In the following discussions, the "it-is-what-it-is" nature of the data sources will be an integral part of our schema.

In our case, both administrative data and big data fit the "it-is-what-it-is" concept. They are not necessarily distinct from one another, and from a CSA's perspective, using them means reusing data that originated outside the agency. UNECE (2011) defines administrative data as "data that is collected by sources external to statistical offices" and "administrative sources are data holdings containing information that is not primarily collected for statistical purposes". This broad definition would also include almost all big data, given the existing different definitions of the phenomena. However, an administrative data source does not have to be "big" nor do big data sources normally have administrative purposes. Usually, the administrative data delivered to the CSAs come from and through the operations of another public organization. This is seldom the case with big data sources; they stem from activities, events and operations within the whole of society.

2.2. *The "Elephant in the Room" – Representativeness*

Unfortunately, there is an elephant in the room when we deal with "it-is-what-it-is" data sets. The elephant is the fact that these source files may not be appropriate for making statistical inferences concerning the general population because the selection probabilities are non-probabilistic. A recent American Association Public Opinion Research (AAPOR) task force report on non-probability sampling (Baker et al. 2013) stated that "approaches lacking a theoretical basis are not appropriate for making statistical inferences". It was pointed out in an earlier AAPOR report (Baker et al. 2010) that statistical estimates and inferences drawn from "it-is-what-it-is" data sets cannot be trusted to be representative of the general population. In reference to the two AAPOR reports, Langer (Langer 2013) quotes a well-known classical reference (Kruskal and Mosteller 1979) stating that "[w]e prefer to exclude non-probability sampling methods from the representative rubric." Langer is implying that one cannot ever claim that results derived from an "it-is-what-it-is" data set are representative of the general population. This is a strong statement and raises questions about the ultimate usefulness of "it-is-what-it-is" data sources.

As the 2013 AAPOR report states, the key is the risk associated with the source data set not being representative of the general population. This is a serious risk because most "it-is-what-it-is" sources suffer from significant coverage issues associated with various sub-populations within a target population. (Overcoverage, duplicates, undercoverage, and missing data can occur in any data source and they can all lead to a population or sample being nonrepresentative. For brevity, at times we will use these terms interchangeably or in a generic sense.) The risk of bias is a systemic problem when dealing with "it-is-what-it-is" data sets and, as illustrated in Subsection 4.4, the cross-linking of "it-is-what-it-is" data sets significantly increases the potential risk. This is the Achilles heel of "it-is-what-it-is" data and, if we cannot address this issue, we will never be

able to widely use "it-is-what-it-is" data. As survey methodologists, we must be able to defend ourselves from criticisms of bias caused by, for example, undercovering populations, such as the underprivileged or rare populations. Without a methodology to measure coverage issues and correct its effects, how do we maintain our credibility? Our schema offers a strategy for confronting this key issue and a stepping stone enabling CSAs to handle a paradigm change and make statistics by repurposing and combining data from sources outside their direct control.

2.3. Correcting Nonrepresentativeness with Registers and Frames

In our article, we address the representativeness risk by creating statistical population registers or frames that allow us to measure and correct over- and undercoverage. Most CSAs estimate for three types of populations: persons, businesses (from a National Accounts perspective these include nonprofit and public organizations) and geography. So, we propose that CSAs adopt these registers as their fundamental mechanism for dealing with nonrepresentativeness within their data ecosystem. As we go through the next few sections we will outline a strategy that:

1. Creates three lighthouse (base) registers systems and uses them to measure under- and overcoverage in various strata. Then, we use these registers to create calibrations (design-based designs) or models (model-based designs) or Bayesian priors (Bayesian designs) to correct for under- and overcoverage.
2. We will assume that we can construct a stratification definition process that ensures that within each stratum we can assume that the observed entities were generated by a random process. Thus, within each stratum the observed entities are assumed to be representative of the stratum sub- population.
3. This estimation capability will be supported by efficient and frequent monitoring of entity transitions in the registers. We foresee the regular use of indicators of entity flows and means of validation (through surveys and investigations) that are regularly used to update the strata and tombstone information in the registers.

The authors recognize that their strategy is naive and pseudo-scientific. It may not correct for all the biases created by the "it-is-what-it-is" data sets' under- and overcoverage. Yet it is a first step along a well-trodden design-based path. As we gather more expertise, future methodologists will develop more mature and complex methodologies for dealing with representativeness. By anchoring "it-is-what-it-is" data over time against better known or controlled data, there will be progress. While we recognize the risks of following a strategy that does not have an unambiguous theory behind it, we feel there is no other choice.

3. Creating the Lighthouse Registers

3.1. Estimation and Representativeness Requirements Imply the Need for Registers

Our evolutionary schema's estimation strategy is built upon three lighthouse registers systems (Thygesen and Grosen-Mielsen 2013). The structural supports for estimation will

be the three traditional entity registers/frames already used by many CSAs. These are geography (or land, dwellings, property, or addresses), persons (or households, or families) and firms (or organizations, businesses, enterprises, establishments, or plants). Traditionally, the census played the role of both the dwelling and person registers, while the business register provided a firm register. To these three key base registers we add time, so that cause and effect relationships can be studied. As in science, time is a special dimension with unique properties quite different from our other three entity dimensions. Yet nevertheless, it will enter many estimation problems. Note that in our schema, a base register system is not equivalent to a sampling frame nor to a census. A census could be an input for building a register and a frame is an output of a register system. In practice, base register systems might be a single database file spanning all time periods and sub-populations or it might be a collection of subregistries achieving the same purpose. Base registers: define important statistical units, define standardized populations, contain links to units in other base registers, contain links to other data sources that relate to the same units, are important as a sampling frame, and can be used for demographic statistics for the units (Wallgren and Wallgren 2014). For convenience, we will dispense with using the descriptor "system" and use the singular form "register" when referring to base register systems.

The base registers of geography (LR), persons (PR) and businesses (BR) together with time are the lighthouses for estimation in their respective dimension. They illuminate potential areas of bias in our estimation system and shine light on the quality of our estimates (Figure 1). Base registers that are connectable to our data sources are the key design element that will allow us to make high quality estimations. These registers will be our starting point for adjusting for nonrepresentativeness effects in our schema.

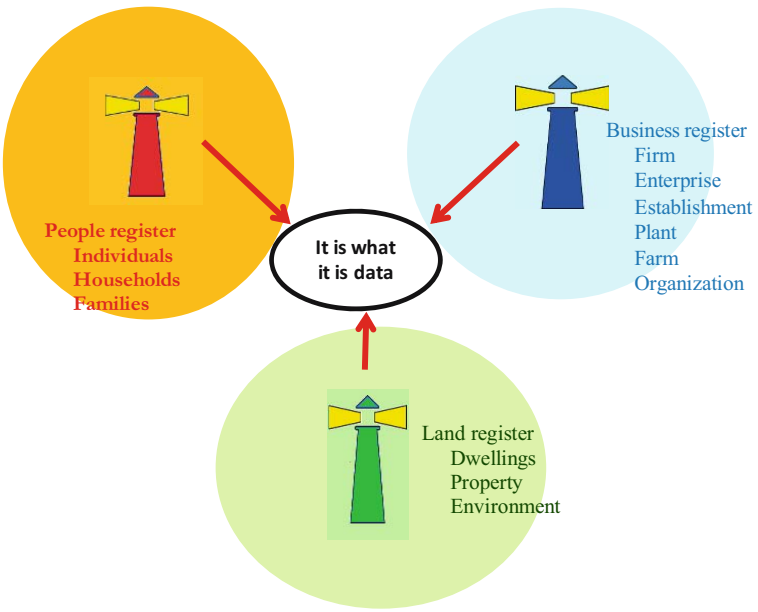

Fig. 1. The three base registers anchor estimates and illuminate the quality of "it-is-what-it-is" data sets.

Historically CSAs have built business registers from administrative files, while censuses and administive address files have been the main sources to make pseudoregisters for dwellings and persons. We envision a future where all three base registers are built from administrative files supplemented by surveys and censuses.

The challenge is how to create a base register from a disparate collection of administrative files. If we create a union data set of all the entities covered by multiple administrative data set sources, the resulting population size is often orders of magnitude larger than the current estimated population count of entities. Alternatively, when we create an intersection data set for the available administrative data sources, the coverage of the population rapidly plummets as more administrative data sets are joined. Typically, the entity count in the intersection data set is much lower than the current estimated population of entities. Winnowing down the number of entities in this collective to create an unbiased frame with complete unduplicated coverage of a desired target is a complex task. We believe that this implies that we must recognize representativeness as the core issue in developing our register systems. The data sources will be "it-is-what-it-is" data sets and we accept that fact and deal with it the best we can. Without some structured strategy for dealing with this issue, any estimation process will be of poor quality. Fortunately, CSAs have a template for developing these future registers: the current universally accepted process for creating business registers.

In the following sections, we will outline a strategy for constructing a representative base register. There are commonalities of function and design that cross the three base registers defined within our schema. Each register will be made up of well-defined entities that exist in the real world and are theoretically finite in number and countable. In each case, the registers will cover a wider population of entities than the current active population. The register will be created from multiple sources of varying quality, indicating whether an entity existed or not and over which period. Entities will have birth and death dates. Hierarchical structures may exist within each register: for businesses (enterprises, establishments, plants, locations); for persons (households, families, persons); and geography (regions, census tracts, land holdings, buildings, addresses).

3.2. Universal Identifiers

Most of the literature on creating person registers was written by authors from countries where universal and unique personal identifiers exist and have been in use by the general population for decades (Bakker and Daas 2012). If a country has a universal identifier given to every resident at birth or upon entering the country for residence, then they have the core of a personal registry. Yet even in this fortunate case, this will not be sufficient information for creating a base person register. To create a PR, one also needs entry and exit dates from the country and birth and death dates for every entity. Alternatively, a country that conducts regular censuses has the core for creating a person register. But in this case, they must coherently merge all the available censuses into a unified file, plus augment the information from other sources detailing births and deaths, and entries and exits to the country. Adding time to our schema complicates our design.

Outside of Europe, universal identifier registers are rare. In most countries, there are serious technical and political obstacles to overcome if one wishes to create a universal

identifier system. The authors believe that for the foreseeable future, most of the countries in the world will not have a universal person identifier system. Therefore, a generic register construction methodology should not depend on the existence of a universal identifier. This requirement presents us with a conundrum if neither a universal identifier system nor regular censuses exists. We are confident that the current available BR technology, together with an evolutionary development strategy can overcome this challenge. In the following sections, we will assume that a universal identifier does not exist.

3.3. A Template for a Base Register

The basic BR template is mature, well understood and supported by a broad international consensus. Thus, we will use a simplified BR structure as a template for illustrating how one might construct and maintain the three base registers. In common usage, the term "business register" is not a base register system in the sense that we define it. The current standard design of the business register encompasses a complex system of interacting files, rather than one core list. In our schema, we define three types of files encompassed within each base register system: the source files or *administrative registers;* the *entity register* which is our core base register; and the *statistical registers,* which we might think of as statistical frames.

The BR's *entity register* of firms enumerates all business entities that have an event in any "it-is-what-it-is" source data files in the past Ω years. These collections of source files used to identify and birth entities to the entity register we will call the *administrative registers* and are subregisters within the business register system. The BR's administrative registers can give conflicting or partial information concerning the presence of business activity at any point in time. They can have different processing dates with different lags in their arrival at a CSA. Different data source agencies tend to use different identifiers. Duplicate transactions can occur. The information collected by each source can be radically different, with conflicting evidence concerning events and activity. Thus, the entities birthed from the administrative registers into the *entity register* represent a spectrum of entities with a varying quality of information. Some entities will have definitive birth dates and continuous ongoing economic activity over a span of time. Others may only show evidence of an entity registration, with no sign of subsequent activity. We divide this spectrum into three groups. The groupings depend on the quality of the administrative information indicating whether the entity exists and is active in a specific target population at a specific time. At the top of the spectrum are entities with multiple confirmed indicators of existence from high quality sources and at the bottom of the spectrum are firms with only partial information from one low quality source.

All sources are not considered of equal quality or informative value concerning the presence of activity. As such, rules must be created that define when the activity observed implies an actual birth to the register. When maintaining the register, there is a trade-off concerning the breadth of information included in the entity register versus the cost of processing and maintenance of this information. Typically, a small number of sources are viewed as "fundamental" indicators. New registrations from these fundamental sources will always generate births to the base register. Internal IDs for these fundamental sources

are maintained within the entity register. Typically, these maintenance processes ensure that the internal IDs are consistent over time and space, and are unduplicated. If there are multiple IDs on the registers, there should exist an internally generated unique ID that spans the population of all the fundamental sources. If hierarchical structures exist on the register, then multiple internal IDs may be required. The entity register can include "inactive" entities: entities that show uncertain, infrequent, very weak or no signal of activity. (For business entities, "inactive" is an acceptable terminology, but for the PR it would be inappropriate. A preferred terminology then might be a classification of "unconfirmed".) An example of an inactive firm is a single indicator of registration or creation on a single administrative file with no known other event. Figure 2 summarizes this process.

The full register system is not useable in any real sense by users outside CSAs because it contains a collage of disparate populations and plethora of active and inactive entities. The essential process in Phase 1 is the "birthing rules" shown in Figure 2. In the BR, these rules are typically set by a cooperative team from the BR, National Accounts, business surveys and methodology. External users of the register cannot change or control these rules.

The administrative and entity registers are never seen by external users. Instead, the users see the *statistical registers* (or target population frames). A frame is the empirically derived list of the target population of interest. In the design-based paradigm, it would be our sampling frame and most practitioners think of a frame in the context of a sampling but, it is also our best possible estimated entity count. The register system presents statistical registers (or survey frames) to users by putting a filter over the full register. Each filter changes what entities the user will see. This is the second phase of the register maintenance system. Figure 3 below summarizes this process. The key process in phase 2 is the "statistical register creation rules".

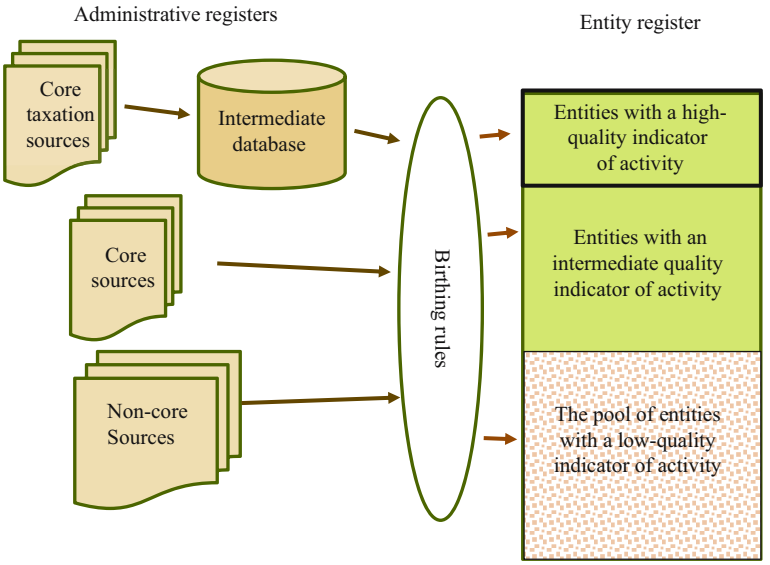

Fig. 2. Phase 1: The birthing and maintenance processes for generating an entity register.

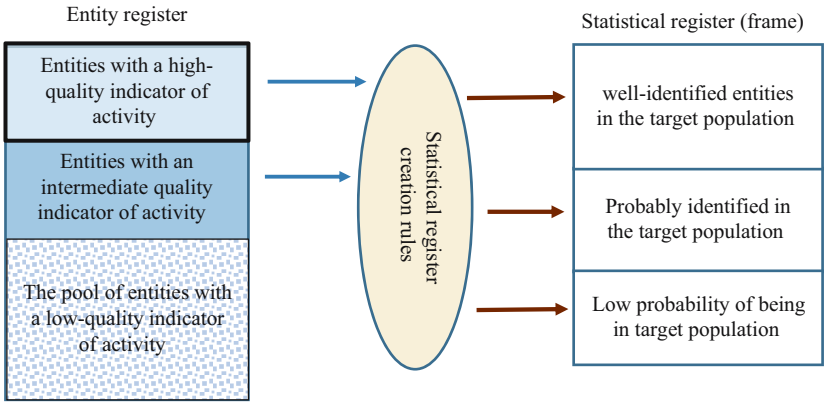

Fig. 3. Phase 2: Creation of statistical register and target population frame.

The real output of the registers is frame lists from the BR, PR, and LR of target populations at a given time for example, lists of active firms, resident populations or property holdings. In our paradigm, it is these *statistical registers* that become our lighthouses for identifying and decreasing potential biases.

The "statistical register creation rules" are a filtering mask that uses standardized rules to extract statistical registers. Users will have no control over the definitions of the standardized filtering fields, but they will have considerable flexibility within these definitions. In the case of the BR, the user may extract any standard industrial code (SIC) grouping, but they will not be able to change SIC definitions. Similarly, BR users would not be permitted to change the definition of "inactive firms" but they could choose to select active and/or inactive firms. Of course, inactive firms will not be maintained to the same standards as active firms.

Typically, a CSA will generate a full business frame (statistical business register) every publication period (monthly, quarterly or annually). Thus, a design principle of a statistical business register is that it must be possible to recreate the business frame that was used for a particular publication date. Meanwhile, updates keep flowing into the BR and influencing the view of these historical periods. In general, while revised historical frames could be created, the original frames are used instead. This would be an appropriate strategy for the other base registers. It is information from the time stamped versions of the Statistical Register that we suggest should be used to calibrate estimates and make the results from using "it-is-what-it-is" data sets more representative (less biased) of the target populations. We will discuss this further in Section 5.

In a base register system, there may be a third phase, a feed-back loop between the estimation processes from the evolutionary databases and the register processing. For example, in the context of the BR, business survey responses can lead to updates of name, address, industry classification, and so on. In the case of our evolutionary data bases environment, the estimates that come from the various integrated data sources are analyzed and the knowledge discovered can be fed back to the entity register. These feedback loops are important for keeping the core registers up-to-date and as accurate as possible. We assume the updates will be tombstone information at the entity level.

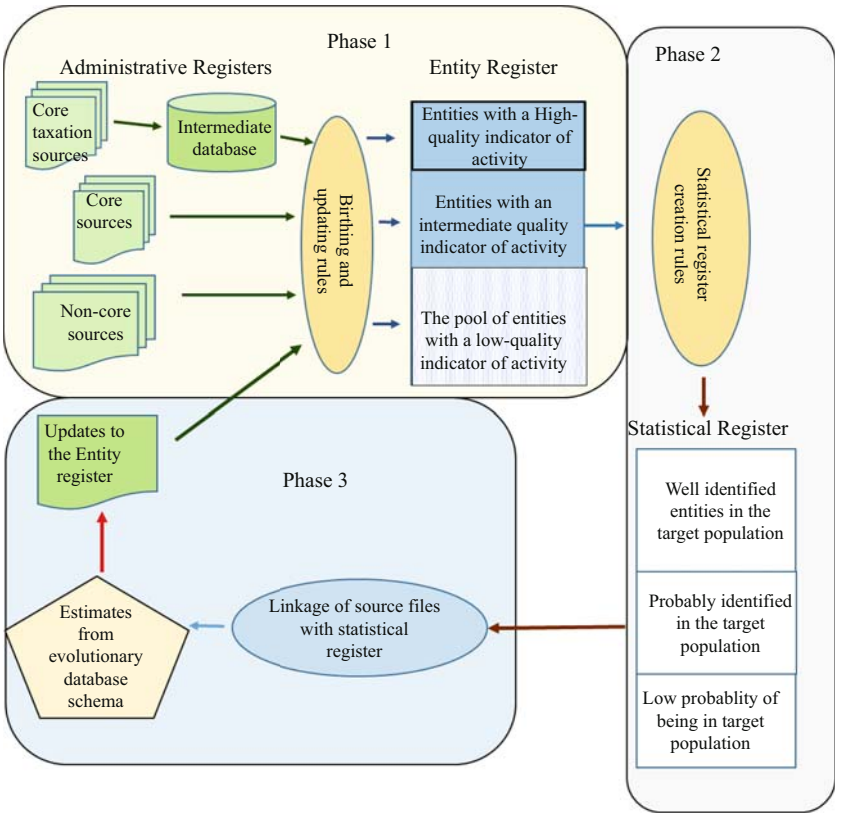

Fig. 4. Putting it all together: a base register system.

Event information is not maintained in the register. Observed shortcomings in the register's design may also lead to feedback. This can be particularly important in a build-up stage to monitor and tune the rules for "birthing" and "statistical register creation" in phases 1 and 2. As the CSAs become more proficient at generating these intermediate files, they will provide the registers with a continuous maintenance function. Figure 4 illustrates all three phases as one integrated register maintenance system.

4. The Evolutionary Database Schema – The Data Model

4.1. Time and Cause and Effect Relationship

CSAs tend to view time as a descriptive characteristic rather than a fundamental dimension. Time becomes an estimation domain much like sex, age, race, and so on. Yet, observations are events in time and when we combine two or more data sources we need to know how to time order the events observed in the data sources. Many social scientists intuitively grasp this point because they are looking for cause and effect relationships or they wish to understand how social systems are evolving. For social scientists, time is a transcendental variable that helps them make sense of estimates. CSAs tend to think in

terms of cross-sectional estimates (or panels) in time rather than a time series evolution. Time series analysis questions are often "end of the line" analyses that marginally affect the cross-sectional survey designs.

Time opens avenues for us to use, analyze and improve the quality of our data sets. Observing related events (a timeline of events) for an entity can provide us with a sense of the evolutionary changes in our data or the volatility of measurements over time. This can provide us with proxies for measuring quality. Having a timeline of events for a common individual allows us to develop improved methods for detecting and fixing errors that are localized to one time period. When one wishes to link entities and events in disparate data sets, the time lines can provide extra information that can improve the quality of the linkages and in some cases, it may allow us to develop quality measurement tools for the cross-linkages.

Time in our evolutionary schema is a fundamental concept and every recorded event must have a time stamp. There is a time-ordering of all events in the schema, so we can distinguish between events, such as diagnosis, treatments and results. Time can open new avenues to improve editing, linking and quality measurements. For an interesting and expanded discussion of the importance of time in statistical analysis one might read Dunn (1946).

4.2. Event and Timeline Databases

To illustrate how the evolutionary time schema might work, let us consider an example of a researcher who wishes to test whether a causal link exists between wealth later in life and education. The data sources available are two administrative data sets, an annual filing of income tax returns, and a collection of school records from a group of school boards. We can view each of these data sets as a list of unique entities and associate a set of date-stamped events with each entity. Each entity's record can be viewed as a timeline of observed events for that entity. Because new events will be constantly added to the database timelines, the timelines are always evolving in time. The events are containers holding the information gathered for this event. In practice, the information might be just a date stamp and virtual pointers to a record in a subsidiary database.

4.2.1. The Annual Filing of Personal Income Tax Returns

Let us call each collection of files that come from a common generating mechanism and contain common frame entities and identifiers a "timeline database". Thus, the collective of all the taxation filings by individuals through time would be the timelines database of an individual's annual tax filings. The data within this database would be structured in a specific manner. The fundamental unique key in the database might be the individual's taxation number and an individual's annual tax filing would be an event in that individual's timeline. Note that there is no requirement that the data be collected on a fixed periodicity and entities' annual filings could be missing in some years. Within each timeline database, the collection of entities must be of a common type, but different timeline databases could contain different types of entities or events. Using survey terminology, there must be a common frame unit within each timeline database. In our schema, the grouping of events based upon a common entity ID within a single timeline

Entity i	**Event 1**	**Event 2**		**Event n**
Entity ID	Event date	Event date		Event date
characteristic 1	Event type	Event type	…	Event type
characteristic 2	characteristic 1	characteristic 1		characteristic 1
	characteristic 2	characteristic 2		characteristic 2
measurement m	measurement k	measurement k		measurement k

Fig. 5. One entity's timeline in the taxation timeline database.

database is viewed as a deterministic function and not a linking process which we view as a nondeterministic process. Entity IDs are assumed to be known true facts. In our terminology the process of creating the entity timeline in Figure 5 will be referred to as a "grouping" function.

Maintaining the taxation database could be straight forward and cost effective. Whenever a new batch of annual filing comes in, one only needs to find the associated Entity ID in the database and add a new event to the record. If the evolutionary database consisted of containers with virtual pointers, one would only need to update the pointers. If the database was structured properly, this activity might require minimal re-indexing and sorting. Once a new batch of records was appended to the end of the timelines database, it might never be touched again. The imposition of the timelines schema onto the taxation database does not require the owners and previous users of the subsidiary tax databases to change any of their previous methods. If these databases remain static, no changes will occur in the timeline database. Even edits of the subsidiary data sets may not require any updates to the timeline database. Only additions, deletions or changes involving Entity IDs will require a recompilation of the groupings inside the event containers. For an expanded discussion on the evolutionary data base structure, one might read Chapter 3 in Lothian et al. (2017).

4.2.2. The Collection of Education Events from a School Board

For the education timeline database, the unique identifier might be a student while, the events might be results from tests, special education evaluations, discipline reports, and so on. Each event would occur at a specific time and there would be characteristics defined for each event. Some characteristics might be defined at the identifier level (like birth date, last address, name), while others might be defined at the event level (like date of transaction, observed attribute, and so on). In this timeline database, multiple different types of events might be recorded; each being pulled from a different subsidiary database. Again, the records must all relate to a common entity and there must be a mechanism for grouping a student's events into a timeline.

Identifier inconsistencies could be a significant issue with educational data. Changing schools could lead to the generation of an alternate student number and home address. In addition, as children age they can change their desired names. These types of inconsistencies might result in students being assigned multiple identifiers and causing fragmentation of the student's event timelines.

The database design and programming would be evolutionary, so that mechanisms can be developed at future dates to resolve these inconsistencies. The long-term solution for this issue is to make the unique key an internally generated database ID that can be remapped to join up the fragments. For an expanded discussion on this issue, one might read Appendix B in Lothian et al. (2017).

The two timeline databases examples were chosen to illustrate the proposed timeline database structure and give a flavor of the challenges that CSAs would face when implementing this structure. In our design, we emphasize flexibility and evolution to deal with the constant changes in the number of data sources and what is contained within each source. By "farming out" the control of the data sources we are implicitly accepting that each source is an "it-is-what-it-is" data source. Note that timeline databases are not lighthouse registers, but the lighthouse registers might use information from timeline databases.

4.3. Linking or Relating Timeline Databases

One of the intended purposes of the evolutionary schema is to allow users to explore cause and effect relationships and to cross-relate disparate "it-what-it-is" data sets. To accomplish this task, users of this evolutionary schema will want to build and discover linkage relationships between different timeline databases. As an example, they may want to explore the relationship between school performance and health and wealth in three data sources. To build the output data set, one must start with one of the timeline databases and connect to a second timeline database through a linkage relation database. Then one must connect the resultant amalgamated database to the third source database using a second linkage relation database. Figure 6 illustrates how the potential relationships (linkages)

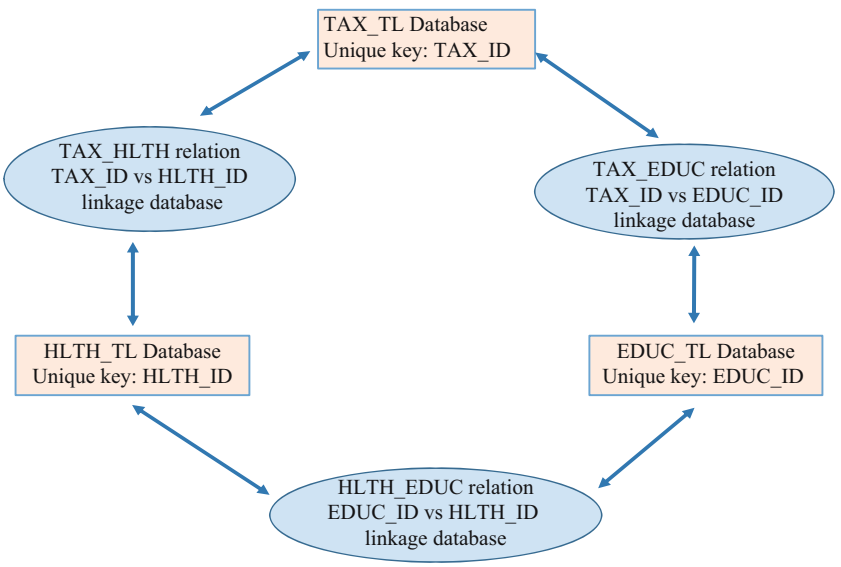

Fig. 6. Functional relationships in the evolutionary system.

between the timeline databases are defined. The ovals in Figure 6 will be referred interchangeably as linkage functions or linkage databases.

The linkage databases are the tools that will allow us to cross-relate timeline databases and explore causal models. The linkage databases will contain unique key pairs defining relationships (links) between two timeline databases. All linkages are assumed to be one-to-one and are not necessarily exhaustive. Only direct relationships can exist between two timeline databases. Linkages involving three or more timeline databases only exist indirectly and the solutions are path dependent. Thus, if A, B, and C are three timeline databases then $(A \cap B) \cap C \neq (A \cap C) \cap B$. The intersection of the three databases is path dependent and is neither commutative nor transitive. Anyone who has observed Google searches is aware that the results of the search depend on the order of the words used to do the search. This can lead to inconsistencies in the produced results, but a generalized multi-path linkage function is not feasible. CSAs will have to establish orders of precedence for multi-source linkages.

Our database design is evolutionary in many senses. The algorithms linking, editing or transforming the data will incrementally evolve as more knowledge is acquired. Initially, some linkage relationships might be undefined and implemented on a need-to-have basis. Fields in subsidiary databases, events, timelines and new survey data sources can be iteratively added as they become available or are needed. The maintenance of the evolutionary database will be devolved and distributed amongst local groups with strengths and experience in the local data. The timeline databases can be disseminated among unrelated control groups and separate teams could be assigned responsibility for creating and maintaining the linkage databases. Each team could add events; edit fields independent of the other groups. New linkage technologies could be implemented without requiring any revisions to the timeline databases or their subsidiary databases. It becomes a distributed and cooperative evolutionary system where local changes will not force a recompile of the complete system.

4.4. Cross-Linking Disparate Data Sets Significantly Increases Risks

Intuitively, most humans have a sense of the Law of Large Numbers. We know that small observation sets are untrustworthy. So, it is natural to assume that adding more data to our system must make it more trustworthy, but that is not true when cross-linking timeline data sources. For linked data sets, this premise does not hold. To demonstrate the problem, let us return to the example we used in Figure 6, where we are linking health and wealth data sources, so we can explore how health affects wealth. Let us make a few conceivable assumptions. We will assume that each data set is drawn for the same population, but suffers from coverage issues. Perhaps, the TAX data does not cover some of the people who never entered the labor market and thus earns no revenue and the health data only partially covers children. Figure 7 illustrates what happens when you link the data. The integrated data set (small oval) acquires the weaknesses of both source data sets. Integrating additional data set sources only makes things worse. When dealing with integrated data sets, one should always assume the output linked data set will not be representative of the target population.

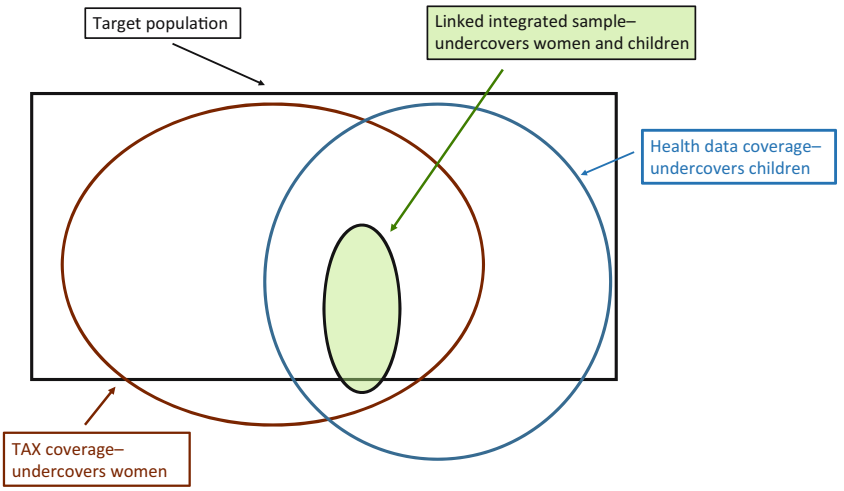

Fig. 7. Coverage biases in integrated data sets.

Representativeness is the central issue when dealing with it-is-what-it-is data sets and this section illustrates that however we attack the problem, there are no magic bullets. Instead, with representativeness in mind, one must use a structured and methodical strategy to identify and eliminate biases related to coverage problems. Our approach transfers the representative problem to the lighthouse registers in Figure 1. There dedicated teams can focus on continually identifying and improving representativeness in an evolutionary manner. Like in the case of the BR, new sources or strategies will be found to minimize coverage biases.

4.5. Timelines are a Fundamental Concept

In the data model presented, time is a foundational concept. Our intent was to design a database that could relate causes and effects, and a necessary requirement for this is ordering events from multiple timeline databases into a single event timeline. The linking of events into a timeline can open avenues for improved linkage strategies. Missing linkage variables can be estimated from other events in the timeline and inconsistencies in names, addresses, age, and so on, can be edited and standardized by analyzing the full timeline. The linkage strategy could depend on time vectors instead of a single value. Perhaps this kind of linkage strategy could help us deal more effectively with name and address changes.

4.6. Linkage Processes have an Underlying Probabilistic Nature

Linkage functions are assumed to be probabilistic, in the sense that the linkage function always has significant uncertainty associated with each identified link. Linkage functions produce a subsample of the two source data sets, where the records produced have an underlying probabilistic element. The number of links found and the "truth" of each link is probabilistic (random) in some sense. While it is almost certainly true that the linked data

set is generated by some probabilistic process, we have little knowledge concerning the selection probabilities. We are not even certain whether the selection is with or without replacement. A key issue is whether the probabilistic sample is representative of our target population. What we will assume is that the sample selected will be non-confounded (at random) (Rancourt et al. 1994) within some estimation domain. This is a powerful assumption.

Our linked output data set is a convenience sample (i.e. non-probabilistic samples sometimes referred to as opportunity or accidental samples see for example Baker et al. 2013) which we will refer to as a linked sample. Our linked sample is a subpopulation derived using relationships to which we have access, but we have no control or knowledge of how these relationships were constructed. Researchers using this linked sample cannot make scientific generalizations about the general population from this sample because it may not be representative of the target population. Strictly speaking, linked or convenience samples are non-probability samples (Baker et al. 2010 and 2013), yet we can hypothesize a hidden underlying probabilistic selection mechanism that is random within some estimation domain (stratum). Thus, the non-probabilistic element of the selection process only affects the balance between estimation domains. The credibility of a researcher's results when using this hypothesis will depend on convincing the reader that the researcher has properly compensated for the imbalance between domains and that the final estimates are representative of the population of interest.

4.7. Using Stratum Definitions to Improve Representativeness

CSAs create various types of strata to help address nonrepresentative issues and to improve the efficiency of estimates. Coverage issues are regularly encountered in CSA surveys and censuses. Even full-enumeration censuses can experience significant undercoverage of special subpopulations. CSAs have several strategies for dealing with these types of issues. One standard practice is to assume that nonresponse is missing at random within a stratum. This is analogous to the strategy that we are proposing.

CSA methodologists are mindful of the potential weaknesses caused by assuming that nonresponse is missing at random. Yet, nonresponse and undercoverage of special subpopulations occurs in every survey and census. This has forced CSA methodologists to develop a toolbox of strategies to eliminate coverage issues. These may include: assuming missing at random within strata; modelling using auxiliary information; Bayesian imputation; targeted follow-up surveys of under-represented subpopulations; calibration; capture-recapture techniques; propensity models; and so on.

Historically, CSAs have followed an evolutionary strategy in developing these methodologies. The authors see a comparable evolutionary development strategy occurring within our schema. Our assumption of "random selection within a stratum" is a first step in this development chain. We expect that, over time, more sophisticated technologies for dealing with coverage issues will arise.

Our lighthouse registers and assumption concerning randomness within a stratum are initial building blocks that force methodologists to confront representative issues head on. What we are proposing is a heuristic strategy based on what worked in the past, rather than

a theory built from first principles. If a CSA can construct the three lighthouses and if the statistical registers are a reasonable approximation of complete unduplicated coverage of the target populations, then the authors believe that toolboxes for fixing representativeness can be developed for "it-is-what-it-is" data sets.

5. Estimation Plays a Central Role in the Design of the Evolutionary Schema

Up to this point, the focus of our database design was implementing a cause and effect design and exploiting the scalability, flexibility and efficiencies of an evolutionary data model. The objective is to use estimates derived from this schema to make inferences concerning real world populations and how entities in these populations interact in time and space. (For a further discussion on the importance of space and supporting GIS solutions in our paradigm see Lothian et al. 2017).

To derive interpretable estimates from the data, we must make some conjectures about the data and apply a structured scientific estimation theory. From survey theory, we suggest borrowing from one of the three different paradigms. The first is the design-based, or randomization, theory (Särndal et al. 1992), which emphasizes that attribute values of the records in the data are fixed values and that it is the random selection of the elements in the data set that ensures the representativeness of the target population through the use of a sampling frame. The second paradigm is the predictive, or model-based, approach (Valliant et al. 2000), where the values are regarded as realizations of random variables and the design by which the survey elements are selected is of less importance. A third paradigm that can be implemented is a Bayesian inference framework. It has been put forward as useful for analysis of small non-probabilistic samples and appears to be a direct alternative when data from different sources are being combined (Little 2012 and 2015; Rao 2011). The choice of which paradigm to use depends both on one's estimation objective and/or philosophical training. In our schema, we present a design-based paradigm, but either of the other two paradigms could be substituted. Our schema does not favor any of the three paradigms. One just needs to choose one paradigm and stick with it.

In the literature to date, there has been much discussion on the data availability, potential data models, building the databases and linking algorithms, rather than how estimation and statistical inference enters the data schema. Most of the literature seems to focus on database structures or building the information technology infrastructure (Holman et al. 2008). There are numerous articles on linkage algorithms (Fellegi and Sunter 1969; Jabine and Scheuren 1985; Winkler 2009); others on specific attempts to define the required data structures (Holmberg et al. 2011; Wallgren and Wallgren 2014); and others on the construction of specific data ecosystems (Holman et al. 1999). Literature with an end-to-end perspective of all the components necessary to do statistical inference and estimation are less common. One such overview is the paper by Zhang (Zhang 2012), which provides a conceptual statistical methodological framework for using "it-is-what-it-is" data sets (or administrative data sets in his article). This article was inspired by Zhang's article. Our schema provides one possible implementation framework within Zhang's conceptual model.

Discussions on estimation strategies are complex and the non-probabilistic nature of the "it-is-what-it-is" data sources can generate considerable controversy. As we mentioned

previously, the major obstacle is constructing an estimation framework that generates results that are representative of the general population. We believe that registers must play a central role in making estimation representative. Registers and frames are the support scaffolding for estimation done within the evolutionary schema. Without this scaffolding, our estimation strategies will be weak and prone to failure irrespective of the design paradigm that we choose to use.

5.1. Registers/Frames Anchor the Evolutionary Databases

Generally, statistical frames will be derived from information available from one of the three base statistical entity registers using the Statistical Register Creation Rules (Figure 3). If U^R is the set of all entities in our base register, then:

$$U^F \subset U^R \tag{1}$$

The base or entity register U^R is an integrated, micro-merged and maintained list of entities created from the combination of different administrative data sources based on identifiers that are unique to the various data sources, (see Section 3). In addition, we need a linkage relationship database that cross-links the statistical frame and the entity timeline database of interest. With U^F defined, we could calibrate the "it-is-what-it-is" linked sample to the frame. (In Bayesian terminology, these calibrations would be prior constraints on the probability distributions.) By using aggregate data from the frame, such as domain totals X_d and domain counts N_d and regarding them as known constants, weights and calibration equations can be constructed that reproduce these known parameters within the linked data set. Hence, we construct weights w_k that satisfy the principal expressions:

$$\begin{aligned} X_d &= \sum_{A_d} \omega_{d,k} x_{d,k} = \sum_{U_d^F} x_{d,k} \\ N_d &= \sum_{A_d} \omega_{d,k} = \sum_{U_d^F} 1_{d,k} \end{aligned} \tag{2}$$

where A_d is the subset of linked elements k observed within domain (stratum) d. U_d^F is the complete enumeration of elements k within domain d in the survey frame F while $\mathbf{x}_{d,k}$ is a vector of known variables provided by the frame. Depending on the circumstances, different variants of Equation (2) can be used (Särndal et al. 1992).

In our formalism, we assume that U^F provides unduplicated complete coverage of the target population U^T. To cover domains d, we just add the subscript d and define variables

$$x_{d,k} = \begin{cases} x_k & for \quad k \in U_d^F \\ 0 & for \quad k \in U^F - U_d^F \end{cases} \tag{3}$$

where an important case of x_k are the indicators 1_k variables that gives us, $N_d = \sum_{U_d^F} 1_{d,k}$. The estimation of other parameters, for example, an unknown total $Y_d = \sum_{U_d^F} y_{d,k}$ from the linked data set is then done by using the same set of calibration weights, that is, $\hat{Y}_d = \sum_{A_d} \omega_{d,k} y_{d,k}$.

This is a popular technique often used in a design-based inference (Lundström and Särndal 2005; Särndal 2007). To apply it here, we have to make the naïve assumption that the linked sample within an estimation domain is a nonconfounded sample from the target population. Then calibrated estimates are possible, and under this simple scenario one could estimate variances. (Nonconfounded might be considered as synonymous with "observations missing at random". We assume that the observed subpopulation is representative of the full target population in every variable. Nonconfounded has a wider contextual meaning that implies the measurement variable is unbiased in both the statistical and non-statistical sense. The term "unconfouded" is discussed by Rancourt et al. 1994).

If we have entity level information from the frame, we can take the calibration technique one step further and apply explicit models, that is, compute calibration weights that, instead of Y_d and N_d produce model predictions of super-population parameters Y_d^* and N_d^* (Wu and Sitter 2001). In this case, a separate model can be applied for every study variable and domain, although that would require a considerable modelling effort.

In a Bayesian framework, we can denote the information we have from the statistical entity frame by Z, it would enter in the specification of the prior distribution of the population values, $p(Y|Z)$. It would also be used as covariates during the generation of the posterior distribution and parameter estimation when the prior is confronted with the linked data (Bryant and Graham 2015).

The above discussion is a framework for an estimation strategy rather than an explicit methodology. In practice, we are advocating reproducing classical (design-, model-, or Bayesian-based) estimation techniques used in current survey designs.

5.2. *Linking Disparate Timeline Source Files to the Frame*

Ancillary register/frame information is necessary for unbiased estimation when one is cross-linking two "it-is-what-it-is" data sets. To illustrate this point, we will demonstrate how estimation might occur in a simple example. Let us look at the Section 4 example with our timeline databases of health and income. We wish to estimate the relationships between health and wealth of individual entities in the current population (see Figure 7).

There are multiple difficulties with these data sets. First, the TAX and HLTH timeline databases contain out-of-scope units and have significant undercoverage of the target population. In survey terminology terms, we are neither sure of the true target population size N nor of the true sample size n. Without some knowledge of the true n and N, how can we make unbiased estimates of the relationships? How do we answer such questions as: how many children are expected to have a specific health issue? At best, we can estimate ratio or proportional effects that apply to unknown subpopulations. Second, we know that both timeline databases are confounded, possibly in different ways. In general, the poor, immigrants, the very young, the very old, persons with handicaps, stay-at-home parents, and so on may be missing from one or both timeline databases. In survey language terminology, our linked subsample is a biased sample. However, if we assume that the sample within each estimation domain d is "at random" or nonconfounded, and if the entity frame is of good quality, giving us n_d and N_d we can apply the calibration technique in Equation (2) and improve the estimates by decreasing the bias. Without the register/frame we would not have this possibility.

5.2.1. Linking the Integrated Sample to the Frame

There are two simple strategies one might use to link the integrated sample to the empirical frame. If we are fortunate enough to have quality information available to link the sampled records at the entity level, we can create a micro-entity estimation file with weights applied at the entity level. Even if entity level linkage is not possible, we could link at the domain estimate levels if each timeline database file contains the domain stratum identification variable. In these simple cases, we will be making the naïve assumption that any over- or undercoverage or nonresponse is "at-random" and within an estimation domain is ignorable. As mentioned in Subsection 4.7, CSA methodologists have developed a toolbox of strategies to eliminate coverage issues. We are presenting one of the most straightforward strategies.

5.2.2. Estimation when Entity Level Linkage with the Frame Exists

Let us assume that we have a statistical register of current residents (perhaps derived from a recent census or as a result of maintaining a PR lighthouse). Furthermore, a linkage relationship exists, connecting the resident ID on the frame to the person's personal TAX_ID in the taxation timeline database.

Each ellipsoid in Figure 8 could be considered indicative of a survey processing step or a linkage-step and we call the steps through the linkages 'phases'. Figure 8 illustrates what a phase means in our schema: the pentagon is our output integrated data set from the first phase and comprises two container fields holding the TAX_ID and the HLTH_ID. Without context, the output data set is not generally sufficient information to derive reasonable estimates. In phase 2, the context (or calibration) turns the phase 1 output file into estimates. At times, the second contextual phase can be overlooked when dealing with "it-is-what-it-is data". We believe this is a key point, and frames and registers can provide the scaffolding that supports quality estimates.

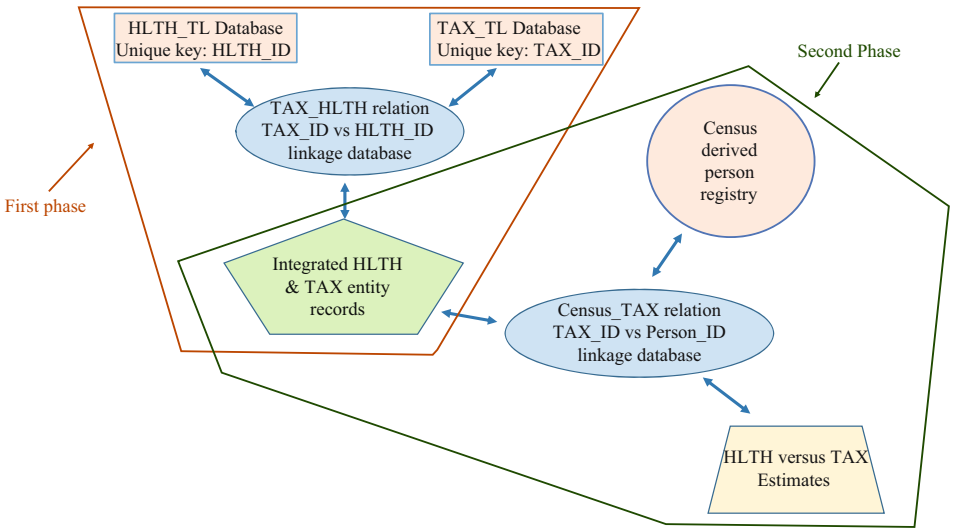

Fig. 8. Estimation processing steps when entity level linkages exist.

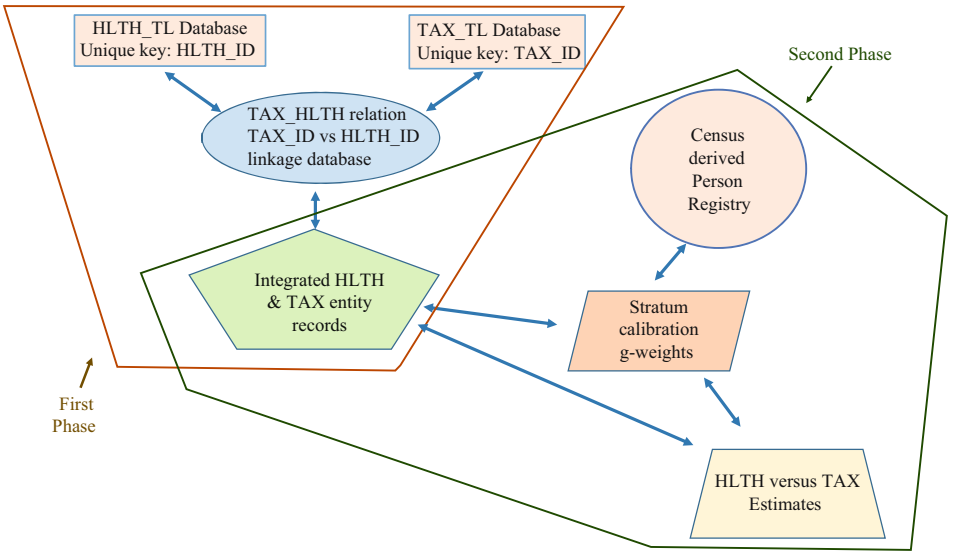

Fig. 9. Estimation processing steps when only common stratum information is available.

5.2.3. Estimation when Entity Level Linkage with the Frame Does Not Exist

In some cases, no reliable entity level linkage information may be available, or it is possible that we wish to link disparate information collected from different, but similar, entities. If the three data sets (health, wealth, census frame) are nonconfounded within an estimation domain, are concurrent and have common domain stratification variables, then an estimator may be found. Figure 9 illustrates this case.

In Figure 9, the linkage between the entity level data (pentagon) and the frame (the circle) will be indirect. The second phase parallelogram represents the combined information from the two. As an example, assume the linked timeline databases contain categorical data such as a geographical (domain d with D categories) and a socio-economic classification (domain d' with D' categories) and that we have this information in the frame as well, but not necessarily simultaneously. With ς being the domain category indicator, we can form $\mathbf{x}_k = (\varsigma_{1k}, \ldots, \varsigma_{dk}, \ldots, \varsigma_{Dk}, \varsigma_{1'k}, \ldots, \varsigma_{d'k}, \ldots, \varsigma_{D'k})^T$ and the right-hand side of (2) will be the domain counts $N_{d\bullet}$ for $d = 1, \ldots, D$ and $N_{\bullet d'}$ for $d' = 1, \ldots, D'$. Hence, we use the marginal distribution of the domain categories from the frame to support the estimation. The weights that satisfy a calibration equation will depend on the inferential principle used and with this $\mathbf{x}_k$ there is no nice expression, however numerical computations are not difficult. In a generic sense, Equation (2) still holds.

6. Total Survey Error Measurement in the Evolutionary Schema

6.1. Error Measurement and Metadata

The authors see information on error measurement as a fundamental component of the evolutionary schema. We see this information being stored and maintained in ancillary metadata files within the ecosystem. We see a metadata file attached to every register,

source data and linkage function file in the ecosystem. The metadata file will contain information on data sources, variables available, discussions on weaknesses and strength of the data and other quality-related information. We see these files as a vital component of ensuring representativeness and long-term quality of estimates.

6.2. *Metadata is the Gateway to the Evolutionary Schema*

Researchers will often approach the evolutionary database with an ambiguous research objective. While these research questions may appear to be wide-ranging or imprecise, they often have very restrictive underlying constraints that will impact on estimation, such as, requiring specific data years and/or subpopulations and/or source data sets and/or relationships being estimated. Researchers will want to know if the available sources/relationships/registers/linkage functions adhere to these constraints. Thus, associated with every object/function in the evolutionary schema should be a metadata descriptor.

The most basic and most requested metadata is an explanation of how to access the data and how it is structured. Users wish to get on with using the data. They require database access protocols, file names and locations, field names, formats and brief descriptions, and source providers for the various data sources.

While this access information is vitally important to users, it presents dangers if it is not balanced with information related to the quality and limitations of the data. Without some discussion on the populations covered by each file, you are encouraging users to apply their analysis to inappropriate subpopulations. Reid et al. (2017) propose a framework for quality documentation and communication in this situation. Or, perhaps, the user is linking two sources with slightly different entities (family versus head of household), and this creates misleading relationships. The authors believe that the priority of meta-data should be the presentation of quality information to the database user, rather than a focus on metadata that is easy to create or requested most often.

6.3. *Measuring TSE*

Total Survey Error (TSE) (see, for example Biemer 2010; Groves and Lyberg 2011) can rarely be reduced to one number, instead it is a structured methodological approach for reviewing and compiling sources of error in a survey. Errors can arise at each classical survey processing step: frame creation, sample design, questionnaire design, questionnaire distribution, collection, editing, follow-up, imputation and estimation. Each survey processing step can introduce errors in potentially different dimensions of error. The TSE paradigm treats each aspect or dimension of the survey processing system as part of a collage that defines the overall measure of quality. At each processing step, quality measurables are collected and assembled into an overall package that allows the survey designer to understand where errors occur and give them some sense of their overall impact on the TSE. TSE is a structured approach to cataloguing and measuring the errors that arise in each of the classical survey processing steps. TSE challenges us to view survey errors in a structured and holistic manner.

While the classical survey processing steps may not be applicable in a data integration paradigm, we believe one should use a similar structured approach that breaks down the

overall estimation process into self-contained subprocesses. A few simple measurables will be suggested within each subprocess. These measures will be naïve, but we believe they will address the primary concerns of most users. We propose focusing on the coverage or representativeness of the specific output data set in the specific subprocess.

6.4. *Error Arising from Source or "it-is-what-it-is" Data Sets*

In our schema, the source data sets are the rectangles in Figures 8 and 9. Source data sets will be administrative, survey or census data sets containing observed variables that we wish to interrelate. Often, the incoming quality and content of these sources will be beyond the control of the CSAs. In the following section, quality measurables will be suggested for these source data sets. Note that the timeline concept introduces a new way to view quality in the schema.

6.4.1. Coverage Statistics

In the Evolutionary Schema, population coverage is a critical concept because it is expected that most data sources will have biases in their coverage of the target population. Every source data set should be related to its coverage of a frame or register, preferably either the BR or the PR or the land register. The population coverage should be as devolved as possible. Thus, for person entity data sources one should provide coverage by sex, age, geography and as many other demographic characteristics as possible.

Thus, for each processing step p and domain d, we will calculate the quality measure set $Q_d^{p,R} = \langle N_d, n_d, f_d \rangle_{p,R}$ where N_d is the size of the target population in domain d of frame U^F which in turn was derived from the base register U^R. Let n_d will be the number of entities in domain d observed in the output data set. Finally, coverage for process p within domain d will be

$$f_d = \frac{n_d}{N_d} \tag{4}$$

If there are m domains in processing step p, the full set of preliminary quality measures $Q_d^{p,R}$ will be:

$$Q^{p,R} = \left\{ Q_{d_1}^{p,R}, Q_{d_2}^{p,R}, Q_{d_3,\bullet,\bullet,}^{p,R}, Q_{d_m}^{p,R} \right\} \tag{5}$$

By examining $Q^{p,R}$ for all d, the analyst/user can derive some sense of the representativeness of the output data set derived from process p. These three measures are simple, and far from comprehensive, yet nevertheless powerful. They address the principle weakness of linked and "it-is-what-it-is" data sets, representativeness. Small values of n_d and f_d or disparate values of f_d for different d can be indicators of quality problems.

We propose using at least one of the three statistical registers/frames to generate $Q^{p,R}$ for each processing step and then placing this information in the metadata. We recognize that these sets of measures are not comprehensive, but they are relatively simple to auto-generate and if they are used properly they will stop the worst cases of misuse of the data. From our perspective, this is the first step in the evolutionary development of TSE indicators of representativeness.

6.4.2. Stability Over Time

These basic measures could be augmented by measuring changes within event timelines. Timelines are collections of events for a specific entity from a common data source. For each entity's timeline, we could record statistics concerning how often a field changes or is missing. Then, aggregate (domain) estimates of the average changes or proportions of missing values could be automatically generated and placed in the metadata. Grouping events into timelines gives us an extra dimension of quality to measure.

6.4.3. Evolution will Develop New Measures of Quality

As new registers, sources, linkage functions, timelines and events get added to the database, users will develop new insights about the data and better measures of quality. We will discover new algorithms to calculate these measures and place them into the metadata.

7. Summary

We set out to identify some key issues in using administrative data: estimation and assurances of quality. In a two-year-long discussion amongst ourselves and other methodologists, we explored the numerous pitfalls one encounters when using administrative data and we discussed several strategies that needed to be a part of any statistical system using administrative data. While we quickly realized that representativeness was the Achilles heel of administrative data, we were strongly influenced by Zhang's article calling for a new conceptual paradigm when dealing with administrative data. Thus, right from the beginning of our discussions we attempted to tackle the problem in a holistic manner, attempting to use a full conceptual paradigm for dealing with administrative and it-is-what-it-is data. Below is a summary of our major finding and an outline of the main features of our schema.

7.1. Administrative Data is Nonrepresentative

The key weakness of administrative data is that various sections of the targeted population have coverage issues, and this generates representativeness problems. Coverage is never 100% in any administrative data source and in most cases, significant portions of a population are under- or overcovered. This is a systemic problem that is considerably worsened by cross-linking multiple administrative data sets. Methodologists must address this key fact steadfastly. We must be capable of defending our estimates from criticisms of bias caused by undercovering the under-priviledged or the rare populations. If we cannot do this effectively then CSAs will lose credibility.

7.2. Correction with Registers and Frames

Our solution to this problem and recommendations to CSAs is to create the three lighthouse registers and use them to measure under- and overcoverage in various domains/strata/classes. Then we use these registers to create calibrations (design-based designs), models (model-based designs), or Bayesian priors (Bayesian designs) to correct

for coverage issues. Here we are following the historical development path for correcting coverage issues in censuses.

7.3. Evolutionary (System Grows in Every Sense Over Time)

We see our schema as an evolutionary system in every sense. New data sources will evolve, and old ones will disappear. New data points will be added (both in time and cross-sections). New database designs will be incorporated, and new estimation and linkage algorithms will constantly be developed. New methodologies will be constantly under development and evaluated.

7.4. Distributed and Collaborative System

Administrative data spans wide areas of knowledge, subject-matter areas, geography and time. In addition, the final design will incorporate ongoing development of complex statistical and IT methodologies. No one group can do these tasks centrally. The tasks and data sources must be delegated across various teams with varying backgrounds and expertise. A central design and control structure would be created to oversee these teams.

7.5. Evolutionary Convergence

When we create our registers, data sources, methodologies, and so on, there must be a path of convergence towards an ever-improving system. Ideally, each new evolutionary step will incorporate all previous information gathered. Consider the BR. In general, most of the information in the BR is tombstone information that rarely changes over time. Births and deaths are a small percentage of the population, only a small percent of the addresses or names change each period, and so on. The BR team focuses on changes rather than the full population. This is also the manner in which the census address list is maintained. Similarly, our evolutionary system would be built along paths that evolve towards better quality and optimality.

Feedback loops are important in this system. Users must be able to feed corrections that they have identified in the registers, algorithms, and so on, back into the system. This is the way the BR and the address register of traditional censuses worked.

7.6. Timelines (Cause and Effect)

There is one message we are hearing continually from researchers who want to use CSA data. They want to do cause and effect studies or longitudinal studies. Our data need to have a natural time structure built in from the beginning. We see each administrative data transaction as an event that occurs at a specific time. We propose grouping and time ordering these events that occurred for a common entity into timelines. For each data source, an implicit timeline database would be created for each entity. Viewing the data in this manner not only allows for effective studies, it also opens new possibilities for editing, imputation, linking, and so on. Of course, there are still statistical challenges in establishing representativeness in longitudinally linked records in order to reliably interpret the results.

7.7. Total Survey Error

We propose a simple strategy for creating measurement tools for quality estimation and we address the representativeness side of the Total Survey Error (TSE) model. In each stratum/class/domain defined by the three lighthouses, we propose calculating the coverage ratio of that data source versus the estimated lighthouse population. When cross-linkages and integrations are done, the stratum coverage ratios should be calculated for the linked data set. This TSE information should be stored in the metadata.

8. References

Baker, R., S.J. Blumberg, J.M. Brick, M.P. Couper, M. Courtright, M. Dennis, D. Dillman, M.R. Frankel, P. Garland, R.M. Groves, C. Kennedy, J. Krosnick, P.J. Lavrakas, S. Lee, M. Link, L. Piekarski, K. Rao, R.K. Thomas, and D. Zahs. 2010. "AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4): 711–781. Doi: https://doi.org/10.1093/poq/nfq048 (accessed May 2018).

Baker, R., J.M. Brick, N.A. Bates, M.P. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90–143. Doi: https://doi.org/10.1093/jssam/smt008 (accessed May 2018).

Bakker, B.F.M. and P.J.H. Daas. 2012. "Methodological Challenges of Register-based Research." *Statistica Neerlandica* 66(1): 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x (accessed: May 2018).

Biemer, P.P. 2010. "Total Survey Error: Design, Implementation, and Evalutaion." *Public Opinion Quarterly* 74(5): 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058 (accessed May 2018).

Bryant, J.R. and P. Graham. 2015. "A Bayesian Approach to Population Estimation with Administrative Data." *Journal of Official Statistics* 31(3): 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028 (accessed May 2018).

Dunn, H.L. 1946. "Record Linkage." *American Journal of Public Health* 36(12): 1412–1416. Doi: http://dx.doi.org/10.2105/AJPH.36.12.1412. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1624512/ (accessed May 2018).

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64(328): 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049 (accessed May 2018).

Ferrara, A., A. Nikolov, and F. Scharffe. 2011. "Data Linking for the Semantic Web." *International Journal on Semantic Web & Information Systems* 7(3): 46–76. Doi: http://dx.doi.org/10.4018/jswis.2011070103 (accessed May 2018).

Fowler, M. and P. Sadalage. 2003. Evolutionary Database Design. Available at: http://martinfowler.com/articles/evodb.html (accessed May 2018).

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065 (accessed May 2018).

Hand, D.J. 2018. "Statistical Challenges of Administrative and Transaction Data." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181(Part 3): 1–24. Doi: http://dx.doi.org/10.1111/rssa.12315 (accessed May 2018).

Holman, C.D., A.J. Bass, D.L. Rosman, M.B. Smith, J.B. Semmens, and F.J. Glasson. 2008. "A Decade of Data Linkage in Western Australia: Strategic Design, Applications and Benefits of the WA Data Linkage System." *Australian Health Review* 32(4): 766–777. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18980573 (accessed May 2018).

Holman, C.D., A.J. Bass, I.L. Rouse, and M.S.T. Hobbs. 1999. "Population-based Linkage of Health Records in Western Australia: Development of a Health Services Research Linked Database." *Australian and New Zealand Journal of Public Health* 23(5): 453–459. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10575763 (accessed May 2018).

Holmberg, A., K. Blomqvist, J. Engdahl, H. Irebäck, L.-G. Lundell, and J. Svensson. 2011. *A Strategy to Improve the Register System to Store, Share and Access Data and its Connections to a Generic Statistical Information Model (GSIM)*. Paper presented at the Work Session on Statistical Data Editing, UNECE, Ljubljana, Slovenia, May 9–11. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp. 37.e.pdf (accessed May 2018).

Holt, T. 2000. "The Future for Official Statistics." *Journal of the Operational Research Society* 51(9): 1010–1019. Doi: http://dx.doi.org/10.1057/palgrave.jors.2600999. Available at: http://www.jstor.org/stable/254222 (accessed May 2018).

Jabine, T.B. and F.J. Scheuren. 1985. "Goals for Statistical Uses of Administrative Records: The Next 10 Years." *Journal of Business & Economic Statistics* 3(4): 380–391. Doi: http://dx.doi.org/10.2307/1391725 (accessed May 2018).

Kruskal, W. and F. Mosteller. 1979. "Representative Sampling, II: Scientific Literature, Excluding Statistics." *International Statistical Review/Revue Internationale de Statistique* 47(2): 111–127. Doi: http://dx.doi.org/10.2307/1402564. Available at: http://www.jstor.org/stable/1402564 (accessed May 2018).

Langer, G. 2013. "Comment: Summary Report Of The AAPOR Task Force On Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1: 130–136. Doi: http://dx.doi.org/10.1093/jssam/smt008 (accessed May 2018).

Little, R.J.A. 2012. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28(3): 309–334. Available at: http://www.jos.nu/Articles/abstract.asp?article=283309 (accessed May 2018).

Little, R.J. 2015. "Calibrated Bayes, an Inferential Paradigm for Official Statistics in the Era of Big Data." *Statistical Journal of the IAOS* 31: 555–563. Doi: http://dx.doi.org/10.3233/SJI-150944 (accessed May 2018).

Lohr, S.L., V. Hsu, and J.M. Montaquila. 2015. *Using Classification and Regression Trees to Model Survey Nonresponse*. Paper presented at the Joint Statistical Meeting (Section on Survey Research Methods), Seattle, Washington, United States. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2015/files/234054.pdf (accessed May 2018).

Lothian, J., A. Holmberg, and A. Seyb. 2017. Linking Administrative Data: An Evolutionary Schema. Available at: SAO/NASA Astrophysics Data System ArXiv. (arXiv:1712.085522 [stat.ME]), accessed May 2018, from Cornell University Library, Available at: http://adsabs.harvard.edu/abs/2017arXiv171208522L (accessed May 2018).

Lundström, S. and S. Särndal. 2005. *Estimation in Surveys with Nonresponse*. Chichester, United Kingdom: John Wiley & Sons, Ltd.

Rancourt, É., H. Lee, and C.-E. Särndal. 1994. "Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Responses." *Survey Methodology* 20(2): 137–147. Available at: http://www.statcan.gc.ca/pub/12-001-x/1994002/article/14423-eng.pdf (accessed May 2018).

Rao, J.N.K. 2011. "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal." *Statistical Science* 26(2): 240–256. Doi: http://dx.doi.org/10.1214/10-STS346. Available at: http://www.jstor.org/stable/23059987 (accessed May 2018).

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics* 33(2): 477–511. Doi: http://dx.doi.org/10.1515/JOS-2017-0023 (accessed May 2018).

Särndal, C.E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33(2): 99–119. Available at: http://www5.statcan.gc.ca/olc-cel/olc.action?objId=12-001-X200700210488&objType=47&lang=en&limit=0 (accessed May 2018).

Särndal, C-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Thygesen, L. and M. Grosen-Mielsen. 2013. "How to Fulfil User Needs – from Industrial Production of Statistics to Production of Knowledge." *Statistical Journal of the IAOS* 29: 301–313. Doi: http://dx.doi.org/10.3233/SJI-130784 Available at: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00784 (accessed May 2018).

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

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics: Statistical Methods for Administrative Data* (2nd edition). Chichester, West Sussex, England: John Wiley & Sons, Ltd.

Winkler, W.E. 2009. "Chapter 14: Record Linkage." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, Vol. 29A, 351–380. Oxford, United Kingdom: Elsevier B.V.

Wu, C. and R.R. Sitter. 2001. "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data." *Journal of the American Statistical Association* 96(453): 185–193. Doi: http://dx.doi.org/10.1198/016214501750333054 (accessed May 2018).

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-based Statistics and Data Integration." *Statistica Neerlandica* 66(1): 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x (accessed May 2018).

Received August 2017
Revised May 2018
Accepted May 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 167–187, http://dx.doi.org/10.2478/JOS-2019-0008

How Standardized is Occupational Coding? A Comparison of Results from Different Coding Agencies in Germany

Natascha Massing[1], *Martina Wasmer*[1], *Christof Wolf*[1], *and Cornelia Zuell*[1]

As occupational data play a crucial part in many social and economic analyses, information on the reliability of these data and, in particular on the role of coding agencies, is important. Based on our review of previous research, we develop four hypotheses, which we test using occupation-coded data from the German General Social Survey and the field test data from the German Programme for the International Assessment of Adult Competencies. Because the same data were coded by several agencies, their coding results could be directly compared. As the surveys used different instruments, and interviewer training differed, the effects of these factors could also be evaluated.

Our main findings are: the percentage of uncodeable responses is low (1.8–4.9%) but what is classified as "uncodeable" varies between coding agencies. Inter-agency coding reliability is relatively low κ ca. 0.5 at four-digit level, and codings sometimes differ systematically between agencies. The reliability of derived status scores is satisfactory (0.82–0.90). The previously reported negative relationship between answer length and coding reliability could be replicated and effects of interviewer training demonstrated. Finally, we discuss the importance of establishing common coding rules and present recommendations to overcome some of the problems in occupation coding.

Key words: Occupation coding; coding rules; ISCO.

1. Introduction

Occupations are an important outcome of previous life decisions and a determinant of life chances. Therefore, occupation is an important variable in social and economic research; it allows the analysis of labor market processes, social mobility, and status attainment, to mention just a few. Occupation is a complex construct that is difficult to measure and requires categorization. Most surveys aiming at collecting data on occupation include open-ended questions about the respondent's current or last job. Hence, responses are recorded verbatim.

In order to be analyzed, these textual data on occupation must usually be coded, that is, unstandardized texts must be aggregated into pre-defined categorical systems, for example, the International Standard Classification of Occupations (ISCO) or the Standard Occupational Classification (SOC). Researchers often derive socioeconomic status scores, occupational prestige scores or social class positions (Ganzeboom and Treiman 2003, 159–193), or occupational health hazard scales ('t Mannetje and Kromhout 2003) from these occupational codes.

[1] GESIS-Leibniz Institute for the Social Sciences, B2,1, 68159 Mannheim, Germany. Emails: natascha.massing@gesis.org, martina.wasmer@gesis.org, christof.wolf@gesis.org, cornelia.zuell@gesis.org

Because of the great importance of the occupation variable, it is essential to understand the coding process and to assess the quality of coding results. In the present article, we therefore address the following questions: What percentage of responses to open-ended questions about occupation are uncodeable? How reliable are occupational codings? With regard to the latter question we ask, in particular, to what extent coding results are affected by rules implemented by the coding agency ("house effects"). We explore how the application of different coding rules influences measures of socioeconomic status (SES) and occupational prestige. Finally, we analyze the relationship between the length of answers given by respondents and the reliability of coding across different agencies.

In order to answer these questions, we use data from the German General Social Survey (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS) 2010 and the German field test of the Programme for the International Assessment of Adult Competencies (PIAAC) carried out in 2010. Occupational information was coded into the 2008 version of ISCO (ISCO-08) by three agencies for ALLBUS and two agencies for PIAAC.

This article is structured as follows: We begin by briefly reviewing relevant literature on the coding of occupations and discussing key challenges in this field. Based on this presentation, we develop hypotheses that will guide our empirical analysis. After a short description of the data and the coding procedures used in our study, we present our results. We conclude with a discussion of the main findings and implications of our study.

2. Challenges of Occupation Coding and Previous Research

Transferring textual information on occupations into numerical information is a multifaceted and demanding task. Errors can occur at several stages and impact the quality of coding. To date, both theoretical attempts to systematize and explain factors affecting the quality (i.e., validity and reliability) of occupational coding results and empirical research on this topic have been limited. Elias (1997, 13) distinguished between 1) problems relating to the extent and quality of the *data to be coded*, 2) problems relating to the *classification itself,* and 3) problems relating to the formulation and application of *coding rules* and the *coding process*.

In interviewer-administered surveys, interviewers must ask the questions and record the information provided by respondents. The first prerequisite to obtaining valid and reliable information on occupation is the use of adequate questions. The usual recommendation is to use at least two separate questions (International Labour Office (ILO) 2012, 55f.) to obtain both information about the job title and about the main tasks and duties performed in the job. Hoffmann et al. (1995) point out that the questions should be simple, containing familiar, widely understood terms. Furthermore, they underline the importance of specific design features. For example, including instructions and examples in the questions seems to have different effects depending on the education of respondents. They also show that the size of the text fields provided for the answers affects the answer length because it serves as a cue to the level of the detail expected (Hoffmann et al. 1995). Hak and Bernts (1996) point out the fact that the interpretation of answers by the coders is a preliminary key step of the coding process. In particular, when respondents used vague or ambiguous

terms or provided contradictory information, the assignment of a specific occupational code is mainly a matter of interpretation. One can try to improve the quality of the responses by appropriate interviewer training (Billiet and Loosveldt 1988).

Indeed, several authors have reported challenges encountered during the interview process that might result in problems when coding answers to the questions about occupation. For example, Schierholz et al. (2017) reported that respondents tend to provide incomplete or contradictory information, which influenced coding results. Similarly, Geis and Hoffmeyer-Zlotnik (2000, 113) reported that typically 15% to 25% of answers cannot be completely coded to the most detailed level. Therefore, our first hypothesis is:

H1: We expect the percentage of not completely codeable answers to be around 20%.

The quality of coding is not only influenced by the question used or the interviewer training, but also depends on the precision, completeness, and clarity of the coding scheme. In this context, the descriptions of the categories play a crucial role in helping coders to find the correct codes. However, changes in occupational specialization reduce the applicability of coding schemes over time, and the typical lengthy updating phase of such classifications might be too long to capture ongoing changes in occupational structures. Coding schemes developed for cross-national comparison are even more challenging because they have to find the right balance between categories reflecting international comparative structures and national specificities. In this article, we cannot test the effect of the coding scheme itself on the quality of codings, as we use only one coding scheme in our empirical analysis.

Basic criteria reflecting the quality of coding results are inter-coder reliability – that is, the extent to which the same code is assigned to a given text by different coders – and internal validity – that is, the extent to which the most appropriate code is selected. For our data, we do not have any codings that could serve as a "gold standard" to assess validity. Therefore, we will focus on reliability, keeping in mind that high reliability does not guarantee high validity, but certainly is a necessary requirement. Several measures of reliability are available, for example, simple percent agreement, Cohen's kappa, Scott's pi, and Krippendorff's alpha (see, e.g., Freelon 2010). We will report Cohen's kappa – a coefficient measuring agreement adjusting for agreement occurring by chance. With an increasing number of categories, random agreement is negligible and Cohen's kappa (multiplied by 100) is only marginally lower than percent agreement.

One way to increase the reliability of occupational coding is to formulate and apply rules capturing the definitions and general guidelines of an occupational classification (see, for example, Geis and Hoffmeyer-Zlotnik 2000). Usually, the occupational classification manual includes some general rules. For example, in the publication *International Standard Classification of Occupations*: *Structure, Group Definitions and Correspondence Tables,* which is referred to in what follows as the "ISCO manual", ILO (2012) suggests that the following three rules (in that order) should be applied when classifying jobs with a broad range of tasks and duties: (1) If the tasks and duties require different skill levels, the job should be classified in accordance with the tasks and duties that require the *highest level of skills*. (2) If the tasks and duties are connected with different stages of the production and distribution process, tasks and duties related to the

production process should take priority. (3) In cases with tasks and duties at the same skill level and at the same stage of production and distribution, the job should be classified according to the predominant, that is, most time-consuming, tasks performed. Detailed explanations of specific boundaries (for example, concerning the distinction between managers and supervisors or operators of small businesses) are included in the ISCO manual. Also, recommendations are provided regarding the use of job-related information other than job title and main tasks and duties actually performed. The developers of ISCO (ILO) recommend that decisions should be based on the tasks actually performed, rather than on any other information.

However, beyond general clarifications and rules, more specific operational rules are needed to determine the most appropriate occupational code for a given description of a particular job. When automatic dictionary-based coding is applied, the program that assigns codes to text can be considered part of these rules. In addition, to ensure consistency, previous coding decisions will usually be incorporated into agencies' coding rules. Finally, recurring incomplete or ambiguous answers require rules on how to deal with these cases. As an example, Ganzeboom and Treiman (1996, 210) recommend providing coders with information on the numerical sizes of the specific occupational categories, and thus on the probability of category membership, in order to help them code ambiguous cases. Specifications on whether and how further job-related information (e.g., employment status or size of organization) are to be taken into account in case of doubt are also an important part of practical coding rules. This is especially relevant if important information about performed tasks is lacking in the input material. In the actual process of assigning codes to verbatim answers, the individual coders will – more or less systematically – use and "enrich" these rules in their own – more or less idiosyncratic – way. This may be a source of bias, especially when the correlated coder variance is high because of high workload per coder (Campanelli et al. 1997, 444–445).

Belloni et al. (2016) studied coding errors in occupational data in the Netherlands using data from the Survey on Health, Ageing and Retirement in Europe (SHARE). The authors recoded responses to open-ended questions about occupation for the Dutch sample of the SHARE data using software for semi-automatic coding: all cases above the certainty score threshold of 70 were coded automatically; all residual cases were coded manually by an expert coder. This coding was used as a benchmark and compared to the results of the coding according to the standard procedure implemented in SHARE (manual coding by trained coders). Inter-coder agreement at the one-digit ISCO level was 71% for current job and 72% for last job; at the three-digit ISCO level, it was 52% for current job and 56% for last job.

Campanelli et al. (1997) reported results of two similar empirical studies, in which they compared results from different coders of the Office of National Statistics (ONS) and applied manual, computer-assisted and computer-automated coding methods on data of the British Household Panel. Data were coded in accordance to the UK Standard Occupational Classification (SOC), a classification similar to ISCO. Reliability for manual codings was between 0.75 and 0.80. Validity, assessed by comparisons with expert coders, yielded agreement rates, ranging between .69 and .84. Only modest gains in reliability and validity could be obtained by using computer-assisted methods.

Several studies have dealt with the question of how the quality of coding results is influenced by the *rules applied during the coding process*. For manual coding, Hak and Bernts (1996) argued that the effectiveness of training in terms of inter-coder reliability is improved, not only by "communicating coding instructions to coders (*theoretical* training)", but also by "socializing coders into practical rules" beyond the general coding instructions. Such practical coding rules are, at least partly, specific to the coding agency (or individual coders) and, as such, reflect their experience and expertise. Thus, coding rules can have a negative effect, as they can lead to what Bushnell (1998) referred to in a different context as the "coding system bias." Depending on the coding agency and the rules in place, results of coding can differ systematically between agencies ("house effects"). The higher the differences between these agency-specific rules, the higher is the likelihood of house effects in occupational coding, that is, that the coding will differ systematically between agencies. Thus, although these house effects increase the reliability within one agency, they may reduce reliability between different agencies and jeopardize validity. Thus, our second hypothesis is:

H2: The coding reliability between different coding agencies is lower than the coding reliability within agencies.

In social research, occupational codes are often used to calculate occupational prestige and socioeconomic status (SES) scores (Ganzeboom and Treiman 2003). The agreement between prestige and SES scores derived from the codings of different agencies will be higher than the inter-agency agreement between the occupational codes. This is necessarily true because merging a large number of codes into a smaller number of categories results in higher agreements per se. Additionally, we argue that in the cases where the choice between different codes is hard to make, the occupations in question are more similar with respect to prestige and SES than a randomly chosen pair of occupations would be. In a study by Maaz et al. (2009), occupational information was coded by professional coders and lay coders (trained student assistants). The level of consistency between these codings was generally not as high as desirable. For example, for the ten one-digit codes of the International Standard Classification of Occupations of 1988 (ISCO-88), the authors reported a Cohen's kappa (κ) of around 0.67. For the 390 more detailed four-digit ISCO codes, κ did not exceed 0.5. When occupational codes were converted into International Socio-Economic Index (ISEI) scores, correlations were generally higher (between 0.75 and 0.85). Therefore, our third hypothesis is:

H3: Socioeconomic status and occupational prestige scores will show high levels of agreement.

Finally, we are concerned with the relation of answer length and codeability. Previous research suggests that longer text strings do not necessarily result in more reliable codings. Conrad et al. (2016) found that longer descriptions of occupations were less reliably coded than shorter descriptions. For coders, texts that provide too much detail or are too complex can be difficult to interpret within the framework of an occupational classification and can make it difficult to decide on a unique occupational code. The Conrad et al. (2016) study corroborates Bergmann and Joye (2005, 9f) that ". . . the more detailed the information to be sorted into occupational groupings, the less reliably individual cases are assigned to

categories." It is also in line with the findings of Cantor and Esposito (1992). They found that coders who were asked to comment on interviewer recordings of occupations only rarely indicated that these should contain more specific information; some even criticized the fact that interviewers had provided too much information. Thus, our fourth hypothesis is:

H4: Long answers to open-ended questions about occupation do not result in higher coding reliability.

Before we present analyses on our hypotheses, we briefly discuss the data and approach we use in our empirical study.

3. Data, Classification, and Coding Procedure

3.1. Data

To test our hypotheses, we use data from the German General Social Survey (ALLBUS) 2010 and the German field test of the Programme for the International Assessment of Adult Competencies (PIAAC) conducted in 2010. Both ALLBUS and PIAAC were carried out as computer-assisted face-to-face interviews with randomly selected respondents from official population registers.

We restricted the samples to respondents in gainful employment aged between 18 and 65 years. In our analysis, we considered the respondents' current or last occupations and their parents' occupations. We did not differentiate between these different types of occupations because coding reliability showed no substantial differences.

Respondents' job titles and activities were measured with a two-part open-ended question in ALLBUS and a three-part open-ended question in the PIAAC field test (see Table 1). These questions were embedded in different background questionnaires that also

Table 1. Questions about current occupation used in ALLBUS and PIAAC.

ALLBUS	PIAAC
1. What work do you do in your main job? Please describe your work precisely.	1. What job are you in at the moment?
	2. Please describe this job exactly. Please give the exact title of the job. For example, rather than just saying "management assistant" give the full title "management assistant in freight forwarding"; instead of merely stating "worker", give the exact title, for example "machine fitter". If you are a civil servant, please give your exact grade, for example "police sergeant" or "tenured secondary school teacher". And if you are a trainee/apprentice, then state the profession in question.[1]
2. Does this job, this work have a special name?	3. Does this job have a special name?

[1]A similar instruction was provided in ALLBUS in the general interviewer training. However, this instruction did not appear on the screen and was therefore not read out to the respondents.

included other occupation-related questions. The ALLBUS respondents were first asked about their status in employment using a very detailed classification that distinguished white-collar workers according to their tasks, civil servants according to their career paths, blue-collar workers according to their qualifications, employers by number of employees, and farmers by the size of their utilized agricultural area (for more details, see Supplemental data, Table A1). This was followed by the open-ended questions about occupation presented in Table 1. The first sub-question of ALLBUS asked for a description of work done in the job. Strictly speaking, only the second sub-question referred to the key information "job title", although in practice, many respondents reacted to the first sub-question by naming the title of their job. After these questions, several other occupation-related questions (e.g., about supervisory status) were asked. The original question wording in German can be found in the Supplemental data (Table A2).

The question about occupation in PIAAC was similar to, but more detailed than that in ALLBUS (see Table 1). There were three sub-questions about occupation, starting with a sub-question explicitly asking for the exact job title. In contrast to ALLBUS, all sub-questions required an answer (a text entry or a click on the "don't know" or "refused" button). Also in PIAAC, these questions were followed by several occupation-related questions (e.g., the industry that the respondent was working in, and supervisory status).

In both surveys, respondents who were currently not employed, but who had worked in the past, were asked about their last job. Furthermore, respondents were asked about their mothers' and fathers' occupations when the respondents were 15 (ALLBUS) or 16 (PIAAC) years old. The wording was similar to that of the questions about the respondent's occupation, however only limited ancillary information was provided to the coders about the parents (formal qualifications and, in ALLBUS, employment status).

To carry out the survey, the PIAAC interviewers were trained very thoroughly. Their training included instruction on how to retrieve and record information on occupation; they were instructed to ask the respondents in a way that the interviewers actually understood what kind of job the respondents were doing. Furthermore, they also had to practice how to ask the occupation questions. In ALLBUS there was no such training. However, the interviewer instructions for the study (in written form) included explanations and examples to illustrate the level of detail necessary with respect to the open-ended questions on occupation.

3.2. The ISCO-08 Classification Scheme

ISCO-08 is the current version of the International Standard Classification of Occupations curated by the International Labour Organization, a specialized agency of the United Nations. It "provides a system for classifying and aggregating occupational information . . . [and] allows all jobs in the world to be classified into 436 unit groups" (ILO 2012, 3). For the purposes of ISCO-08, a *job* is defined as "a set of tasks and duties performed, or meant to be performed, by one person" (ILO 2012, 11). Jobs with very similar main tasks and duties are aggregated into "occupations". Different occupations form the most detailed level of ISCO, unit groups.

Besides *job*, the second main concept underlying ISCO classification is *skill*, which is defined as "the ability to carry out the tasks and duties of a given job" (ILO 2012, 11). More

specifically, in ISCO, occupations are categorized according to the typically required *skill level* and *skill specialization*. Skill level is differentiated into four ordered groups from low to high educational qualifications and accompanying levels of literacy and numeracy. As can be seen in Table 2, Major Group 9: Elementary Occupations is characterized by Skill Level 1; Major Group 3: Technicians and Associate Professionals by Skill Level 3; Major Group 1: Managers by Skill Levels 3 and 4; Major Group 2: Professionals by Skill Level 4; and all other major groups by Skill Level 2. Major Group 0: Armed Forces Occupations is an exception, in that it can contain a broad range of skill levels.

Within the major groups, occupations are grouped by the type of skill specialization, that is, "the field of knowledge required; the tools and machinery used; the materials worked on or with; and the kinds of goods and services produced" (ILO 2012, 11). It is important to note that the concept of skill level and skill specialization refers to the requirements of jobs and occupations, rather than the skills or education of a specific job incumbent.

Overall, ISCO follows a hierarchical structure where an increasing level of detail is expressed as one- to four-digit codes. As mentioned above, unit groups, which are denoted by four-digit codes, constitute the most detailed level of the classification. These unit groups are aggregated into 130 minor groups, expressed as three-digit codes, which in turn are grouped into 43 sub-major groups, denoted by two-digit codes. At the highest level, the classification comprises 10 major groups (see Table 2).

The structure of the classification is illustrated in Table 3. For example, Unit Group 3112: Civil Engineering Technicians is part of Minor Group 311: Physical and Engineering Science Technicians, which in turn is part of Sub-major Group 31: Science and Engineering Associate Professionals, which belongs to Major Group 3: Technicians and Associate Professionals.

Table 2. Structure of ISCO–08, number of categories, and skill level.

ISCO-08	Major group	Sub-major groups	Minor groups	Unit groups	Skill level
1	Managers	4	11	31	3 + 4
2	Professionals	6	27	92	4
3	Technicians and Associate professionals	5	20	84	3
4	Clerical support workers	4	8	29	2
5	Services and sales workers	4	13	40	2
6	Skilled agricultural, forestry and fishery workers	3	9	18	2
7	Craft and related trades workers	5	14	66	2
8	Plant and machine operators, and assemblers	3	14	40	2
9	Elementary occupations	6	11	33	1
10	Armed forces occupations	3	3	3	1 + 2 + 4
Total		43	130	436	

Source: (ILO 2012, 14, 22).

Table 3. Extract from ISCO-08.

3	Technicians and associate professionals
31	Science and engineering associate professionals
311	Physical and engineering science technicians
3111	Chemical and physical science technicians
3112	Civil engineering technicians
3113	Electrical engineering technicians
. . .	. . .
3119	Physical and engineering science technicians not elsewhere classified
312	Mining, manufacturing and construction supervisors
. . .	. . .
32	Health associate professionals
321	Medical and pharmaceutical technicians
. . .	. . .
324	Veterinary technicians and assistants
3240	Veterinary technicians and assistants
325	Other health associate professionals
. . .	. . .

3.3. The Coding Process

For our study, a number of agencies were contracted to independently code all the occupational data. These agencies are experienced in occupational coding and offer this service commercially; all agencies offer their service roughly in the same price range. All agencies trained their coders according to their own procedures. Most of the coders were experienced in coding occupations, however, we do not have detailed knowledge about their background. The agencies were provided with respondents' answers to open-ended questions about occupation, which had been recorded verbatim by the interviewers. Furthermore, ancillary variables were provided for use by coders during the coding process, including age, gender, formal education, status in employment, public sector employment, supervisory tasks, and industry. Three agencies (labeled here as Agency A, Agency B, and Agency C) coded the responses to the ALLBUS open-ended questions about occupation; two agencies (Agency A and Agency B) coded the PIAAC data. Whereas Agency A was the same agency in both cases, Agency B was not. Agency B refers, in fact, to two agencies: The first agency coded the PIAAC occupation data; its successor coded the ALLBUS data. As the first agency handed over all relevant material to the second agency, and the second agency also adopted the first agency's coding procedures, they are treated as one unit here.

The agencies were not given any further instructions on how to code. Rather, they could apply their typical coding procedures. They could use either semi-automatic or manual coding, and they were allowed to use any coding tools they had available. Furthermore, the agencies themselves decided on the coding strategy to be adopted, that is, 1) the number of coders; 2) the additional material provided (e.g., the complete ISCO manual or sample lists of possible codings); and 3) the organization of the coding process (e.g., coding the major group first and coding the other digits in a second step).

For the coding process, Agency A used a software program that provided the text of the responses. There was no automatic coding involved in the process. However, the coding software offered coding suggestions via templates. These suggestions are based on the

ISCO group labels (e.g., coding digit 1, the suggestions are based on the ISCO major groups, coding digit 4, the suggestions are unit group labels). The main objective of employing this software was to reduce the complexity of the coding process for the coders by using a hierarchical coding strategy, providing the same information to all coders. The coders were not given the ISCO manual as a reference. The coding instructions, ancillary information, conventions, and known problems with suggested solutions were part of an interface in the software.

Agency B used, as a first step, automatic coding based on an extensive dictionary. About 45 to 55% of the answers can usually be coded this way (Geis and Hoffmeyer-Zlotnik 2000, 127). Responses that were not automatically codeable were listed alphabetically, and alphabetically ordered blocks were randomly assigned to coders. In the case of the coding of the PIAAC data by Agency B, an expert coder reviewed the codes assigned, and if systematic errors were observed, coders were given additional training. Several rules were specified for the coders, for example: 1) If two different codes from different major groups are plausible, assign the occupation with a lesser degree of professionalization. 2) If two occupations were mentioned by a respondent, code the first-mentioned occupation. 3) If two different job titles were mentioned, code the more concrete title. The first rule seems to contradict the ILO rule cited earlier, whereby, in the case of several plausible alternatives, the occupation with the highest skill level should be coded. However, we are not sure how "professionalization" was operationalized. To our knowledge, this agency used auxiliary information on employment status in ambiguous cases. We do not know how the two other agencies handled this status variable during coding.

The coding process at Agency C involved two steps: first, occupations were coded according to the five-digit German Classification of Occupations (KldB) 2010 (Bundesagentur für Arbeit 2011). This was initially done automatically, using a dictionary with around 100,000 entries. This coding process was supported by taking into account the ancillary information provided (for example, industry or occupational status). The remaining responses were then manually coded. Several rules were applied, for example: 1) If two occupations were reported, code the first-mentioned occupation. 2) If the open-ended response is inadequate to determine the occupational code, use the ancillary information. 3) If two different codes are equally plausible, assign the code that occurs more frequently in practice. As a second step, the codes of the German Classification of Occupations (KldB) 2010 were automatically mapped to ISCO-08 codes using correspondence tables provided by the German Employment Agency (Bundesagentur für Arbeit 2011).

To control the coding quality within the agencies, and to determine whether low coding reliabilities were due to house effects, we asked the two agencies coding the PIAAC data to code a subsample of answers a second time, assigning them to different coders.

4. Results

In the following sections, we describe our results in the order of our four hypotheses. First, we present results on uncodeable or not completely codeable answers, followed by the description of inter-agency and inter-coder reliability and coding differences. After that, we present consequences for socioeconomic status and prestige scores and finally, we report differences in answer length and reliability.

4.1. Uncodeable or Not Completely Codeable Answers

The number of uncodeable answers was low; to a large extent answers could be coded at the four-digit level (between 78.0% and 97.8%). As can be seen in Table 4, the percentages of answers that could not be assigned to at least a major group by the three agencies that coded the ALLBUS data were 3.4%, 3.0%, and 2.2% respectively. For PIAAC, the respective percentages were 4.9% and 1.8%. These findings are in line with the findings of Hoffmeyer-Zlotnik et al. (2004), who reported 3% of not codeable answers in a survey of the German Environment Agency in 1999. Looking closer at the answers that could not be coded, we found large differences between the agencies: For ALLBUS, 274 answers were classified as uncodeable by one of the agencies, 51 answers were classified as uncodeable by two agencies, but only 28 answers were classified as uncodeable by all three agencies. A clearer pattern could be observed for PIAAC, with Agency A showing a greater tendency to classify a given answer as uncodeable. Agency A could not code 134 of the cases that Agency B coded, whereas Agency B was unable to code only eight cases that Agency A coded. If we look at answers that were coded, but not to the most detailed four-digit level, we find that these sum up to 16% and 19% for ALLBUS and 11% and 9% for PIAAC. These percentages were, thus, lower than we expected (see hypothesis H1). Agencies A and B seem to follow the recommendation of the ILO (2012, 56) to code vague responses to the most detailed level still supported by the information provided. In addition, we found surprisingly large differences with respect to which answers were classified as uncodeable. Here, the different rules of the agencies become obvious. To give an example, one agency classified the response "housewife and typist" as uncodeable (housewife), whereas the others assigned the code for typist (4131). Agency C is a special case in this regard: because they first coded the answers into the Germany Classification of Occupations (KldB) and then recoded these codes to ISCO-08 using the official cross-walk, they used the most probable ISCO unit group when answers were not completely codeable or when the KldB code was not unambiguously transferable.

4.2. Inter-Agency and Inter-Coder Reliability and Coding Differences

For the analysis of the reliability we included all codings of the three agencies. If an agency marked an answer as uncodeable or not codeable on the four-digit level, whereas the other one coded the same answer with a four-digit code, this could be handled as a mismatch in the sense of coding reliability. As can be seen in Table 5, coding agreement

Table 4. Number of codeable digits (in percent).

	ALLBUS			PIAAC	
	Agency A	Agency B	Agency C	Agency A	Agency B
Not codeable	3.4	3.0	2.2	4.9	1.8
1 digit coded	3.7	4.4	0.0	1.4	2.3
2 digits coded	8.2	7.5	0.0	6.0	1.6
3 digits coded	3.7	7.3	0.0	3.4	4.9
4 digits coded	81.0	78.0	97.8	84.3	89.3

ALLBUS: N = 5,130; PIAAC: N = 4,159.

Table 5. Inter-agency reliability (Cohen's kappa).

	ALLBUS			PIAAC
	Agencies C & B	Agencies A & C	Agencies A & B	Agencies A & B
1 digit	0.683	0.685	0.722	0.760
2 digits	0.644	0.632	0.674	0.715
3 digits	0.572	0.528	0.574	0.630
4 digits	0.506	0.475	0.508	0.566

ALLBUS: N = 5,130; PIAAC; N = 4,159.

between the different agencies was far from perfect. We only report κ, because checking the proportion of agreement showed that the differences between these measures are negligible. For ALLBUS, in a comparison of Agency A and Agency B, the inter-agency reliability of assigning an open-ended answer to the same major group (i.e., the same first digit) was 0.72. With $\kappa = 0.68$, the inter-agency coding reliabilities for major groups between Agency A and Agency C and between Agency B and Agency C were even lower. The results are similar for PIAAC: The inter-agency reliability between Agency A and Agency B was around 0.76 for the major groups. As there are only ten major groups, these inter-agency reliabilities are not satisfactory.

Because of the hierarchical structure of ISCO, once the first digit differs, all more detailed codes differ. Thus, reliabilities for sub-major, minor, and unit groups were even lower. In the end, the assigned four-digit ISCO codes differed in nearly half of the cases for ALLBUS ($\kappa = 0.48$ to 0.51), and inter-agency agreement was only slightly higher for PIAAC ($\kappa = 0.57$). This low level of reliability casts serious doubt on the usability of these data for further analysis.

As Table 5 shows, the largest information loss occurred at the level of the major group. Taking the ALLBUS codings as an example, a glance at the distribution of the major groups (Figure 1) reveals systematic differences resulting from the codings of the three agencies involved. In particular, frequencies of Major Groups 3, 5, 8, and 9 differ considerably.

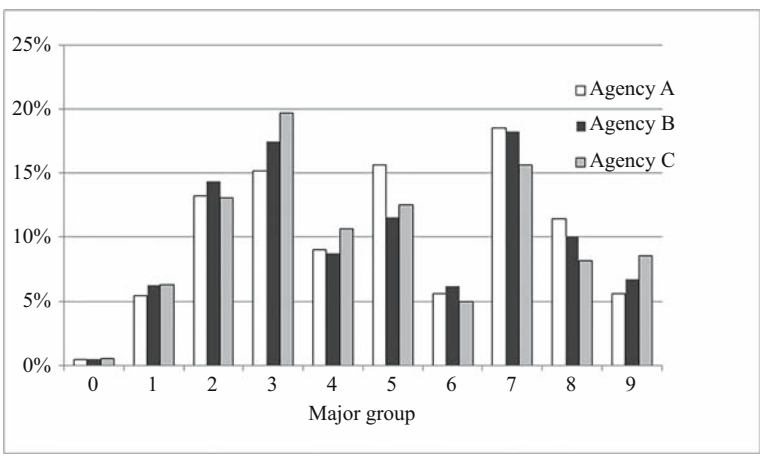

Fig. 1. Percentage distribution of major groups in ALLBUS by coding agency.

The difference is most extreme in Major Group 9, where the number of elementary occupations coded by Agency C was 1.5 times higher than that coded by Agency A.

In what follows, we briefly explore some of the systematic discrepancies we observed in the assignment of the first digit by the different agencies (see supplementary material, Figures A1a and A1b, for relative frequencies of the observed differences on Major Group level; for the most frequent unit group combinations among these first digit discrepancies, see Supplemental data, Table A3). Because it is impossible to list all the differences we found, we concentrate on those differences observed most often and on those that have a large effect on the prestige and SES scores derived from the codes. We identified six such systematic differences:

The first difference related to the handling of self-employed respondents (see Supplemental data, Table A4a). For ALLBUS and PIAAC, Agency B coded a large group of these respondents into Major Group 1: Managers, whereas the other agencies often assigned them to Major Group 5: Services and Sales Workers. To give some examples: if a respondent reported that he or she was a self-employed hairdresser, Agency B classified the occupation into Unit Group 1120: Managing Directors and Chief Executives, whereas the other agencies assigned it to Unit Group 5141: Hairdressers (see Supplemental data, Table A4a). A similar problem was caused by the answer "innkeeper". Agency B classified it into Unit Group 1412: Restaurant Managers; the other agencies assigned it to Sub-major Group 51: Personal Service Workers. ISCO defines "Operators of small cafés, restaurants and bars to whom the management and supervision of staff is not a significant component of the work are classified in Unit Goup 5120 . . ." (ILO 2012, 238). The rules are clearly defined but the answer of the respondents are often not detailed enough to decide what the main task is.

A second frequently occurring difference between the codings for ALLBUS is related to the classification of manual workers. Depending on the agency, these workers were coded either into Major Group 6: Skilled Agricultural, Forestry and Fishery Workers; Major Group 7: Craft and Related Trades Workers; Major Group 8: Plant and Machine Operators and Assemblers; or Major Group 9: Elementary Occupations. Agency C assigned blue-collar workers (e.g., chemical workers, metal workers, or railroad workers) much more often to Major Group 9, whereas, whenever possible, the other two agencies assigned them to a major group other than Major Group 9 (see Supplemental data, Table A4b). Differences arose because many of the responses were not unambiguously codeable. Major Group 6, 7, or 8 would be appropriate if more complex tasks were performed, whereas Major Group 9 should be reserved for those occupations involving only simple and routine tasks.

The third typical systematic coding difference which occurred both in the ALLBUS and the PIAAC codings related to the assignment of responses to Major Groups 3 and 4. Major Group 3 comprises technicians and associate professionals who "perform mostly technical and related tasks connected with research and the application of scientific or artistic concepts and operational methods, and government or business regulations" (ILO 2012, 169). Major Group 4 comprises clerical support workers who "record, organize, store, compute and retrieve information, and perform a number of clerical duties in connection with money-handling operations, travel arrangements, requests for information, and appointments" (ILO 2012, 219). Whereas occupations in Major Group 3 typically require

completion of upper secondary education and possibly a higher education degree (Skill Level 3), most occupations in Major Group 4 require only completion of lower secondary education and possibly vocational training (Skill Level 2). Frequent and typical answers that caused the coding problem in question were "secretary", "clerk" (*Sachbearbeiter*), and "accountant" (see a list of most frequently observed unit group combinations among first-digit discrepancies in Supplemental data, Table A3). To give an example, secretaries may, on the one hand, be what are referred to in ISCO-08 as "administrative and specialized secretaries", who "provide organizational, communication and documentation support services, utilizing specialized knowledge of business activity of the organization in which they are employed" (ILO 2012, 202). In this case, they should be classified into Major Group 3. On the other hand, secretaries may be "secretaries (general)", who "transcribe correspondence and other documents, check and format documents prepared by other staff, deal with incoming and outgoing mail" and so on, in which case they should be classified into Major Group 4 (ILO 2012, 221). Without more detailed information about the respondents' tasks and duties, the occupations cannot be coded with certainty. However, the coding of these answers is of great relevance because of their high frequency and the low inter-agency agreement in relation to these two major groups (around 70%, depending on the coding agencies compared).

The fourth systematic coding difference occurred – again for both surveys – when assigning responses to Major Group 7: Craft and Related Trades Workers, and Major Group 8: Plant and Machine Operators and Assemblers. The main difference between these two major groups is the operation or use of machines at work. Although in the modern industrial world, most crafts require the use of machines, it is not clear from the respondents' answers whether or not the work was done predominantly with or without machines. Thus, the occupation "metal worker" may be classified into Sub-major Group 72: Metal, Machinery and Related Trades Workers or into Minor Group 812: Metal Processing Plant Operators, depending on the stage of development of the industry in the national context.

The fifth systematic coding difference occurred when coding occupations such as educators (*Erzieher*) working in early childhood education and care or youth welfare services. Agency A assigned these occupations mostly to Major Group 5: Services and Sales Workers, Unit Group 5311: Child Care Workers; Agency B assigned them to Major Group 2: Professionals, Unit Group 2342: Early Childhood Educators; and Agency C assigned them to Major Group 3: Technicians and Associate Professionals, Unit Group 3412: Social Work Associate Professionals. At least part of the confusion probably stemmed from the difference between the educational requirements for these occupations in Germany and elsewhere. Whereas in Germany these occupations usually require qualifications corresponding to Skill Level 3, most other countries appear to require Skill Level 4. The ISCO manual is quite clear on this point, stating that "occupations that require the performance of similar tasks should be classified in the same group" and that primary and pre-primary teachers "should all be classified in Major Group 2" (ILO 2012, 28). However, this innovation introduced for ISCO-08 does not seem to have been acknowledged by all coding agencies.

The sixth coding problem was caused by a special national situation in Germany, where – as in some other countries – one type of vocational education and training is "dual

vocational training", which comprises theoretical and practical elements. Most careers in crafts start with such a training program. After successful completion of this program, further training is possible in order to obtain the qualification "Meister" (master craftsperson or engineering technologist). This type of career is very common in Germany. When asked what their occupation was, many respondents answered that they were, for example, a master carpenter or a master electrician. Although no additional information was available on the tasks performed by these master craftspersons, Agency C coded such answers into Unit Group 3122: Manufacturing Supervisors, whereas the other two assigned a code from Major Group 7: Craft and Related Trades Workers that described the craft the respondent was working in. However, the required skill level in Major Group 7 is lower than that in Major Group 3. The reason for the different handling of this type of occupation may result from the coding in KldB and transferring the codes to ISCO: in KldB 2010 all job titles containing the word "Meister" are coded as supervisory tasks, which results in the 3122-coding. Once again, the difference in coding shows that agencies create specific rules to handle national specificities that are not reflected in ISCO. In this case, Agency C applied a standardized rule using the public and official German classification as their standard. As there is no concerted effort to standardize these rules, agency-specific approaches result in different codings and lower inter-agency reliability.

The reliabilities reported in Table 5 reflect not only differences between the agencies' coding rules, but also differences between coders. To disentangle house effects from coder effects, we asked the two agencies that coded the PIAAC data to code a subsample of the occupational information twice, assigning the answers to different coders. We interpret the difference between the inter-coder and inter-agency reliabilities as a house effect that reflects differences in the coding procedures and rules. Inter-coder reliability (κ) of the four-digit codings of our PIAAC subsample was 0.84 between the two coders deployed by Agency B, and between 0.68 and 0.74 at Agency A, where coding was carried out by five coders. As these reliabilities were considerably higher than those between the agencies (0.57), this result supports our assumption that the coding procedures and rules applied by a coding agency affect ISCO coding to a considerable extent. This is in line with hypothesis H2.

4.3. Consequences for Socioeconomic Status and Prestige Scores

As mentioned previously, occupational prestige or SES scores are often derived from ISCO codes. Therefore, we analyzed the extent to which the reliability of these scores was affected by the (comparatively low) reliability of the ISCO codes. Our analysis focused on the International Socio-Economic Index (ISEI) and the Standard International Occupational Prestige Scale (SIOPS), both of which were developed for cross-national research (Ganzeboom and Treiman 2003). Using the syntax proposed by Ganzeboom and Treiman (2012), we calculated these scores separately for each of the coding results of the agencies. In less than 5% of the coded occupations, it was not possible to assign an SES or occupational prestige score. Although the inter-agency reliability of the four-digit ISCO codes was only around 0.5, the consistency between the derived socioeconomic status and prestige scores was much higher than expected. For ISEI, the correlation between each pair of agencies was 0.90, which is higher than the correlations reported in the study by

Maaz et al. (2009). For ALLBUS, the correlation of SIOPS between Agency A and Agency B was 0.84, between Agency B and Agency C it was 0.82, and between Agency A and Agency C it was 0.85. For PIAAC, the correlation between Agency A and Agency B was 0.84. The high correlations of both SES and prestige scores confirm hypothesis H3.

4.4. *Differences in Answer Length and Reliability*

In a final step, we analyzed the length of the answers given to the question about occupation in the two surveys. Because of the differences on how the questions on occupations were phrased in PIAAC and ALLBUS (see Subsection 3.1), we assume that PIAAC answers were more detailed, that is, longer. We measured answer length by counting the total number of words after combining all sub-questions. One drawback of this method is that we were not able to identify word repetitions. Thus, if respondents repeated words, these were counted each time.

Our analysis showed that the questions in PIAAC were answered, in fact, in more detail, leading to answers that were about twice as long as the answers in ALLBUS. The average number of words recorded by interviewers was 2.70 (*SD*: 2.22) for ALLBUS, with a minimum of one and a maximum length of 23 words. For PIAAC, the average number of words recorded was 5.52 (*SD*: 3.87), with a minimum of one and a maximum of 48 words. Furthermore, when systematically checking very common answers, such as "secretary" and "clerk", it became obvious that in PIAAC, as compared to ALLBUS, most respondents had given more details than just a general job title. For example, PIAAC respondents more frequently added information such as "personal assistant to the CEO" to the general job title "secretary", or "office clerk in accounting" to the job title "clerk".

However, the crucial question is whether longer answers result in better – that is, more reliable – codings. To investigate this question, we calculated reliability coefficients based on length of answer (see Table 6). We found that the inter-agency reliability (κ) did indeed vary with the length of the answer. However, contrary to what one might intuitively expect, reliabilities decreased as answers became longer, thus confirming hypothesis H4. This is in line with the finding by Conrad et al. (2016). A closer inspection of long answers (> 15 words) did not yield any specific clues as to the reasons why they appear to be more difficult to code. We suspect that fewer words might result in incomplete occupational

Table 6. Inter-agency reliability (Cohen's kappa) by length of answer.

	ALLBUS				PIAAC	
Answer length	Answers in %	Agencies B & C	Agencies A & C	Agencies A & B	Answers in %	Agencies A & B
1 word	32.1	0.54	0.53	0.54	2.6	0.63
2 words	33.7	0.54	0.53	0.56	22.8	0.61
3 words	9.9	0.49	0.45	0.48	6.7	0.58
4 words	9.3	0.46	0.44	0.47	15.1	0.57
5 words	6.0	0.47	0.37	0.44	13.8	0.57
6 words	3.0	0.39	0.32	0.40	10.1	0.57
> 6 words	6.0	0.37	0.36	0.43	28.9	0.51

ALLBUS: N = 5,130; PIAAC: N = 4,159.

descriptions but that, for frequently used occupational titles, coders have internal coding rules for handling these (partially incomplete) answers. If the answers are longer, coders have more scope for decision-making, and simple rules can no longer be applied. Less consistent coding of longer answers may also be due to the fact that coders are reluctant to read longer texts and therefore base their coding on only parts of the answer.

Although this result is true for both ALLBUS and PIAAC we found, in general, that occupations in PIAAC were more reliably coded than in ALLBUS. For example, the reliability between Agency A and Agency B for one-word ALLBUS answers was 0.54, whereas it was 0.63 for one-word PIAAC answers (Table 6). Interviewer training and a more detailed question seem to have led to longer, more precise answers.

Finally, if longer answers are more difficult to code, it could be assumed that they are also more often uncodeable or not completely codeable. However, our findings do not support this assumption. Uncodeable or incompletely codeable answers were, on average, shorter than codeable answers: the means for completely codeable answers were 2.76, 2.71, and 2.71, while the means for not completely codeable answers were 2.26, 2.39, and 2.30 for Agencies A, B, and C respectively.

5. Conclusion

To better understand how the process of occupational coding affects outcome quality, we studied occupational data from two surveys, each coded by several agencies. We found that the share of uncodeable answers was comparatively small, ranging from 1.8% to 4.9%, depending on the coding agency and data source. However, a closer look revealed that what was deemed "uncodeable" varied between coding agencies. In a next step, we examined the coding results more closely and discovered characteristic, systematic differences in the way agencies interpreted ISCO and defined relationships between answers to be coded and ISCO categories. Notwithstanding the comparatively low agreement between occupational codings from the different agencies, the correlations for SES and occupational prestige scales were satisfactory (between 0.82 and 0.90). Finally, we observe a negative correlation between length of answers (number of words) and coding reliability both in ALLBUS and PIAAC. At the same time, coding reliability between the two agencies (A and B) that coded data from both surveys was higher for PIAAC than for ALLBUS. This difference remained when we controlled for length of answer. In line with Elias (1997, 13) we discuss our findings in relation to the following aspects: the *data to be coded*, the *coding rules and coding process,* and the *classification*.

With regard to the *data to be coded* the seemingly contradictory results concerning answer length are the most interesting. Several factors may have contributed to these findings. First, the intensive interviewer training in PIAAC may not only have increased answer length – with a general negative effect on reliability – but might also have led to "better" answers because interviewers were thoroughly trained on what kind of information was useful for occupational coding. Especially with respect to some frequently reported occupations that are difficult to code without additional information (e.g., educators, secretaries, or clerks) the interviewer training may have had, on balance, a positive effect on reliability. The higher percentage of four-digit codes for PIAAC data also points to the effect of interviewer training.

The second factor that may have contributed to the higher inter-agency reliability of PIAAC codings are differences in the way occupational information was collected. In particular, the first sub-question in PIAAC explicitly asking for the exact job title, may have elicited answers that were more suitable for ISCO coding. In addition, the detailed question about employment status that preceded the open-ended question about occupation in ALLBUS may have played a crucial role. This may have affected the answers given to the open-ended questions because respondents had just given some of this information in the preceding question on their employment status. Besides effects caused by the resulting differences in the core input material for occupational coding, it seems reasonable to assume an interaction effect of data provided and coding rules. The use of this ancillary information for occupation coding seems to have varied in extent or manner between the agencies involved. This could be an important factor that explains the lower inter-agency reliabilities observed for ALLBUS compared to PIAAC, where the question on employment status only differentiated between employed and self-employed.

Taking all these factors together, our results concerning the ALLBUS-PIAAC differences suggest that some of the inter-agency coding disagreements result from various differences in the survey instrument and the interviewer training, and the ensuing differences in the occupational data to be coded – partly in interaction with coding rules applied concerning this data. Therefore, identifying and understanding those characteristics of question wording and interviewer training that lead to the most suitable input material for occupational coding would be a promising field for further research.

Turning to the *coding rules and coding process*, we argue that the systematic deviations that we observed between codings from different agencies indicate that the rules and procedures laid down in the ISCO manual (ILO 2012) do not suffice to completely cover the coding process. This results in low inter-agency reliability and low validity of the results.

From a quality perspective, (inter-coder) reliability is important, but validity of codes is essential. Hak and Bernts (1996) pointed out that the validity of the coding process depends on the quality of coding instructions, whereas inter-coder reliability depends on the implementation of these instructions. This implies that insufficient reliability and validity of occupational codes have common sources. To elicit both reliable and valid information, interviewers should be trained to filter as little information as possible and to ask for an occupational title. Moreover, they should be familiarized with basic coding procedures (Cantor and Esposito 1992, 665).

To improve coding quality, agencies engaged in occupational coding create rules of their own that reflect their interpretation of ISCO, and these house rules lead to the large differences in coding decisions that we found. The shortcomings of the ISCO manual and the secondary, agency-specific coding rules are a serious threat to the validity of occupational codes. We assume that this threat is even larger in the cross-national context where these agency-specific rules most likely reflect particularities of national labor markets. To overcome this situation, we suggest that coding agencies state the rules they apply as clearly as possible and make them public. A publicly accessible body of rules would allow more systematic discussion and development by those engaged in occupational coding. For the time being, survey practitioners commissioning coding services should ask agencies about the coding rules they apply and, where necessary,

negotiate such rules. One may think that the problems resulting from differences in rules may be overcome by applying completely automatic coding routines. Indeed, such methods have been proposed, for example, based on statistical learning (for a discussion of such approaches, see Gweon et al. 2017). However, up to now these approaches are not able to code considerably more than half of the occupations, thus limiting their usability.

With respect to the *classification*, some problems related to coding ISCO originate from the national context in which it is applied, for example, "Meister" (master craftsperson) in Germany. It can be expected that similar problems arise in other national contexts, as ISCO can be seen as a compromise between different national views and peculiarities (see Desrosières 1996, 19–21 for a concise description of the development of international comparative statistics). We have also seen that agencies make fundamentally different decisions when coding occupations. We were particularly surprised to see that there are considerable differences in the use of major groups, pointing to "boundary problems". A further improvement of the ISCO manual, by including even more precise definitions and more detailed explanations, could certainly be helpful in dealing with problematic distinctions.

Given the multitude of jobs actually performed in the real world, rules and procedures laid down in a manual will never suffice to completely cover the coding process. Approaches supporting the coding of ISCO directly in the field may help to avoid some of the problems. The idea behind in-field coding is that respondents have better and more detailed knowledge of the type of work they carry out. Based on additional variables, such as employment status, public sector employment, or supervisory tasks, they could be asked for more details depending on their answer to the first question about occupation. Initial results of such approaches (Tijdens 2014, 2015; Schierholz et al. 2017) are promising. When discussing developments in the way occupational data is assessed and coded, it is also important to look into the effect that new approaches have on the survey process. In-field coding would introduce additional burden on respondents and interviewers, and increase survey costs. Nevertheless, improving results of occupational coding should be a central aim of efforts to improve the overall quality of surveys.

6. References

Belloni, M., A. Brugiavini, E. Meschi, and K. Tijdens. 2016. "Measuring and Detecting Errors in Occupational Coding: An Analysis of SHARE Data." *Journal of Official Statistics* 32(4): 917–945. Doi: http://dx.doi.org/10.1515/JOS-2016-0049.

Bergmann, M.M. and D. Joye. 2005. "Comparing Social Stratification Schemata: CAMSIS, CSP-CH, Goldthorpe, ISCO-88, Treiman, and Wright." *Cambridge Studies in Social Research* 10: 1–35. Available at: https://www.sociology.cam.ac.uk/research/srg/cs10 (accessed January 2019).

Billiet, J. and G. Loosveldt. 1988. "Improvement of the Quality of Responses to Factual Survey Question by Interviewer Training." *Public Opinion Quarterly* 52: 190–211. Doi: http://dx.doi.org/10.1086/269094.

Bundesagentur für Arbeit. 2011. *Klassifikation der Berufe 2010. Systematischer und alphabetischer Teil mit Erläuterungen*. Nürnberg: Bundesagentur für Arbeit.

Bushnell, D. 1998. "An Evaluation of Computer-Assisted Occupation Coding." In *Proceedings of the International Conference New Methods for Survey Research*, August 21–22, 1998: 23–26. Chilworth Manor, Southampton, United Kingdom.

Campanelli, P., K. Thomson, N. Moon, and T. Staples. 1997. "The Quality of Occupational Coding in the United Kingdom." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 437–453. New York: John Wiley & Sons, Inc.

Cantor, D. and J.L. Esposito. 1992. "Evaluating Interviewer Style for Collecting Industry and Occupation Information." *Proceedings of the Section on Survey Methods, American Statistical Association*: 661–666.

Conrad, F.G., M.P. Couper, and J.W. Sakshaug. 2016. "Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes." *Journal of Official Statistics* 32(1): 75–92. Doi: http://dx.doi.org/10.1515/JOS-2016-0003.

Desrosières, Alain. 1996. "Statistical Traditions: An Obstacle to International Comparisons?" In *Cross-National Research Methods in the Social Sciences*, edited by L. Hantrais and S. Mangen, 17–27. New York: Cassel.

Elias, P. 1997. "Occupational Classification (ISCO-88): Concepts, Methods, Reliability, Validity and Cross-National Comparability." *OECD Labour Market and Social Policy Occasional Papers* 20. Doi: http://dx.doi.org/10.1787/304441717388.

Freelon, D.G. 2010. "Recal: Intercoder Reliability Calculation as a Web Service." *International Journal of Internet Science* 5(1): 20–33. Available at: http://www.ijis.net/ijis5_1/ijis5_1_freelon.pdf (accessed January 2019).

Ganzeboom, H.B.G. and D.J. Treiman. 1996. "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations." *Social Science Research* 25: 201–239. Doi: http://dx.doi.org/10.1006/ssre.1996.0010.

Ganzeboom, H.B.G. and D.J. Treiman. 2003. "Three Internationally Standardised Measures for Comparative Research on Occupational Status." In *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables*, edited by J.H.P. Hoffmeyer-Zlotnik and C. Wolf, 159–193. New York: Kluwer Acadamic/Plenum Publishers.

Ganzeboom, H.B.G. and D.J. Treiman. 2012. "International Stratification and Mobility File: Conversion Tools." Amsterdam: Department of Social Research Methodology. Available at: http://www.harryganzeboom.nl/ismf/index.htm. Retrieved 2017/02/27.

Geis, A. and J.H.P. Hoffmeyer-Zlotnik. 2000. "Stand der Berufsvercodung." *ZUMA-Nachrichten* 47: 103–128. Available at: https://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_47.pdf (accessed January 2019).

Gweon, H., M. Schonlau, L. Kaczmirek, M. Blohm, and S. Steiner. 2017. "Three Methods for Occupation Coding Based on Statistical Learning." *Journal of Official Statistics* 33(1): 101–122. Doi: http://dx.doi.org/10.1515/jos-2017-0006.

Hak, T. and T. Bernts. 1996. "Coder Training: Theoretical Training or Practical Socialization?" *Qualitative Sociology* 19(2): 235–257. Doi: http://dx.doi.org/10.1007/BF02393420.

Hoffmann, E., P. Elias, B. Embury, and R. Thomas. 1995. *What Kind of Work Do You Do? Data Collection and Processing Strategies When Measuring "Occupation" for Statistical Surveys and Administrative Records*. Geneva: ILO.

Hoffmeyer-Zlotnik, J.H.P., D. Hess, and A. Geis. 2004. "Computerunterstützte Vercodung der International Standard Classification of Occupations (ISCO-88)." *ZUMA-Nachrichten* 55: 29–52. Available at: https://www.ssoar.info/ssoar/bitstream/handle/document/20762/ssoar-zuma-2004-55-hoffmeyer-zlotnik_et_al-computerunterstutzte_vercodung_der_international_standard.pdf?sequence=1 (accessed January 2016).

International Labour Office (ILO). 2012. *International Standard Classification of Occupations 2008 (ISCO-08): Structure, Group Definitions and Correspondence Tables*. Geneva: ILO.

Maaz, K., U. Trautwein, C. Gresch, O. Lüdtke, and R. Watermann. 2009. "Intercoder-Reliabilität bei der Berufscodierung nach der ISCO-88 und Validität des sozioökonomischen Status." *ZfE* 12: 281–301. Doi: http://dx.doi.org/10.1007/s11618-009-0068-0.

Schierholz, M., M. Gensicke, N. Tschersich, and F. Kreuter. 2017. "Occupation Coding During the Interview." *Journal of the Royal Statistical Society A* 181: 379–407. Doi: http://dx.doi.org/10.1111/rssa.12297.

't Mannetje, A. and H. Kromhout. 2003. "The Use of Occupation and Industry Classifications in General Population Studies." *International Journal of Epidemiology* 32: 419–428. Doi: http://dx.doi.org/10.1093/ije/dyg080.

Tijdens, K. 2014. "Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey." *Journal of Official Statistics* 30(1): 23–43. Doi: http://dx.doi.org/10.2478/jos-2014-0002.

Tijdens, K. 2015. "Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-up Tables." *Survey Insights: Methods from the Field*. Doi: http://dx.doi.org/10.13094/SMIF-2015-00008.

Received August 2017
Revised February 2018
Accepted April 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 189–202, http://dx.doi.org/10.2478/JOS-2019-0009

Modeling a Bridge When Survey Questions Change: Evidence from the Current Population Survey Health Insurance Redesign

Brett O'Hara[1], *Carla Medalia*[1], *and Jerry J. Maples*[1]

Most research on health insurance in the United States uses the Current Population Survey Annual Social and Economic Supplement. However, a recent redesign of the health insurance questions disrupted the historical time trend in 2013. Using data from the American Community Survey, which has a parallel trend in the uninsured rate, we model a bridge estimate of the uninsured rate using the traditional questions. Also, we estimate the effect of changing the questionnaire. We show that the impact of redesigning the survey varies substantially by subgroup. This approach can be used to produce bridge estimates when other questionnaires are redesigned.

Key words: Health insurance; redesigned survey; aggregate model; unit model.

1. Introduction

Health insurance is the primary avenue to receive health care in the United States. People without health coverage are less likely to go to the doctor or hospital than people who have health coverage (O'Hara and Caswell 2013). As such, people with health insurance have higher economic well-being than people who lack health insurance because they have greater access to health care services (Kaestner and Lubotsky 2016). Health insurance coverage also acts as a buffer against the adverse effect of health shocks (McGeary 2009; Bradley et al. 2012), but means that some workers have less job-mobility because of a dependence on their employment-sponsored insurance (Bailey and Chorniy 2016). Furthermore, health insurance is linked to public policy practices and changes in the United States. A key example is the 2010 Patient Protection and Affordable Care Act, the health care law that shaped availability and access to health insurance for millions of Americans.

The Current Population Survey Annual Social and Economic Supplement (CPS) generates widely used estimates on health insurance coverage in the United States, is used to calculate official poverty estimates, and serves as the basis for many policy-related decisions (Blewett and Davern 2006). However, estimates of the uninsured population from the CPS have been historically higher than estimates from other federal surveys (Smith and Medalia 2015). This runs counter to expectation, since the CPS measures health insurance

[1] U.S. Census Bureau, 4600 Silver Hill Rd Washington, D.C. 20233-8500 Maryland 20746, U.S.A. Emails: carla.medalia@census.gov and jerry.j.maples@census.gov.
Acknowledgments: This article is dedicated to the memory of Brett O'Hara. Brett's contribution to the measurement and study of health insurance, medical expenditures, and disability advanced these fields and was of great importance to the U.S. Census Bureau. He was a dedicated mentor to many, and will be greatly missed by his colleagues, family and friends.

coverage in the previous calendar year, while the majority of other surveys measure coverage at the time of interview. When someone is more likely to be uninsured on a particular day than on all days in the year, the calendar year estimate of uninsured persons in the CPS should be *lower* than the uninsured estimates from other surveys, not higher. Research indicated several reasons why this was the case. For example, estimates from the CPS may have actually reflected a mixture of current and past year coverage (Kenney and Lynch 2010), or respondents may have had difficulty with the long recall period (Pascale et al. 2009). Respondents may also be confused about the type of coverage they have. For example, Medicaid coverage has been shown to be misreported as another type of health insurance coverage, but also misreported as being uninsured, which could contribute to an overestimate in the uninsured rate (Call et al. 2008). Another possible explanation focuses on suboptimal imputations of missing data (Davern et al. 2007).

To address these issues, the U.S. Census Bureau implemented a redesign of the health insurance questions in the CPS in 2014, which measured coverage during the 2013 calendar year. One of the major changes implemented in the redesigned survey was improvements to the way data were collected about coverage during the previous calendar year, thereby reducing potential recall bias and clarifying the reference period (Pascale et al. 2016). The redesigned survey asks about health insurance coverage on the day of the interview, and then asks follow-up questions to determine monthly coverage from January 1 of the previous calendar year through the interview date. Additional changes to the questionnaire are explained in Table 1.

The redesigned questions in the CPS lowered estimates of the uninsured rate and brought health insurance estimates more in line with other federal surveys (Smith and Medalia 2015). The timing of this questionnaire change, which completely replaced the traditional questions, was particularly important because it established a strong baseline for measuring health insurance coverage in calendar year 2013, before the implementation of many provisions of the Affordable Care Act. However, at the same time, CPS estimates from prior to 2013 are not comparable to estimates for the period 2013 and beyond, and there is no direct survey-based estimate of the effect of the questionnaire change.

While there are other surveys that measure health insurance coverage that did not undergo questionnaire redesigns during this period, the CPS is unique because it produces the official poverty estimate for the United States; due to the strong association between health insurance and income, it is important to continue the time trend in health insurance coverage in the CPS. Predicting the uninsured rate in 2013 and beyond using the traditional questions makes it possible to continue the historic time series forward, which is necessary to provide an estimate of the effect of changing the questionnaire, and is a central issue when questionnaires are redesigned. This is important because it will allow researchers and policymakers to take a broader view of trends in health insurance coverage from before the ACA's implementation in 2010. In addition, a better understanding of the effect of the CPS questionnaire change provides a measure of the percentage-point difference between the traditional and redesigned estimates for micro-simulation models. Research is needed to derive a model-based bridge estimate between the redesigned and traditional health insurance questionnaires. Our goal is to produce a reliable counterfactual: if the U.S. Census Bureau had kept the traditional health insurance questions in the CPS, what would the uninsured rate have been?

Table 1. Comparison of the traditional and redesigned health insurance questions in the CPS.

	Traditional	Redesigned
Reference period of estimates	Previous calendar year	Previous calendar year
Reference period of questions	Only asks about previous calendar year	Starts with current and then goes back to previous calendar year
Types of coverage	Laundry-list style questions	Starts with general question and then gets more specific
How questions are asked	Collected at the household level	Collected by person; also asks if others in household were covered by fsplan
New content	n/a	Participation in the health insurance marketplace Employer-sponsored insurance offers and take-up Revised medical out-of-pocket expenses

Notes: Health insurance coverage for both is captured at the time of the survey but estimates reflect coverage during the previous calendar year.
For the details on the question wording changes from the traditional to redesigned CPS, see (Pascale 2016).

2. Data

The data for this article come from two sources, the CPS and the American Community Survey (ACS). The CPS is an annual survey of about 98,000 addresses and includes detailed questions regarding health insurance coverage, income received and place of residence. Interviews are conducted from February through April each year, either in person or by phone. We use data from the 2010 to 2013 CPS files, which collected health insurance using the traditional questions about coverage during the previous calendar year. We also use the 2014 to 2015 CPS files, which used the redesigned questions about the months of coverage.

The ACS is a survey of about 3.5 million addresses annually, which collects social, demographic, and housing information. We use the restricted access data that are available through the U.S. Census Bureau's Research Data Centers. The ACS are collected continuously from January to December each year, and interviews are either self-administered (conducted by paper or on the internet), or interviewer-administered (in person and by phone). Note that because there were no changes to interview mode over time, differences in mode between the CPS and the ACS do not affect our results. The health insurance questions in the ACS ask about coverage on the day of the interview.

Both surveys have post-stratified weights and the standard errors are computed using successive difference replication (Fay and Train 1995).

3. Methods

We use two methods to predict what the CPS estimate would have been if the health insurance questions had not been changed: the first is based on yearly aggregates and the second is based on a difference-in-difference model on person-level data. Both methods

rely on the assumption that the time trend in the CPS is parallel to the trend in the ACS over the period 2009 to 2012 (pre-redesign) and 2013 to 2014 (post-redesign), an assumption that we test and validate (Figure 1).

The first approach uses aggregate-level data to predict what the CPS uninsured rate may have been in 2013, had the questionnaire remained the same. To do this, we use the uninsured rate in 2013 from our auxiliary data source, the ACS, plus the difference between the uninsured rates in the CPS and ACS in 2012 (see Equation 1). The difference between the estimates in the CPS and ACS is stable over time; we average the differences between the point estimates of the uninsured rate and the variances between the two surveys over 2009 to 2012 to improve the stability of the estimates. Subsection 6.2., Appendix 2 details the method for calculating the standard errors (Equation A1).

$$Pred\left(Rate_{CPS_{2013}}\right) = Rate_{ACS_{2013}} + \frac{1}{4}\sum_{k=2009}^{2012}\left(Rate_{CPS} - Rate_{ACS}\right)_k \tag{1}$$

This approach has been used in other estimates of health insurance coverage, which used the CPS rate in one year together with the growth rate in other surveys and administrative records (Centers for Medicare and Medicaid Services 2014). As shown in Equation 1, the aggregate approach estimates the uninsured rate primarily, but can also be used to estimate the effect of the questionnaire change secondarily by subtracting the estimated uninsured rate of the traditional CPS in 2013 from the observed redesigned uninsured rate in the same year.

We validate this approach by predicting the uninsured rate in years that we also have observed data, such as 2012, and find that the predicted and observed lines fall on top of each other (not shown). While the aggregate method can provide a good benchmark, it is

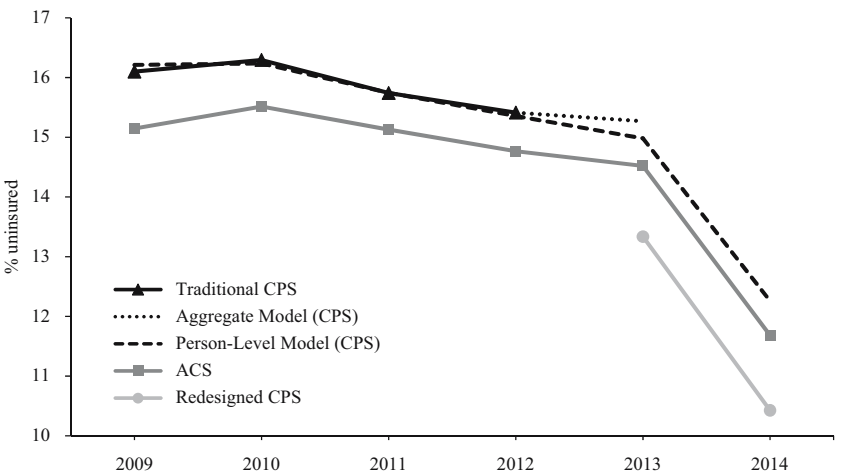

Fig. 1. Time series of the uninsured rate from 2010 to 2014, by data source. Source: 2009–2014 one-year American Community Surveys (restricted data), 2010–2015 Current Population Survey Annual Social and Economic Supplements. Note: Traditional CPS refers to the official uninsured estimate from the CPS for calendar years from 2012 and before. Redesigned CPS refers to the official uninsured estimate from the CPS for calendar years 2013 and beyond.

not practical for the analysis of subgroups due to high variance in the survey. Another limitation to the aggregate model is specific to the case explored here (e.g., health insurance coverage). As many provisions of the Affordable Care Act went into effect in 2014, the relationship between demographic characteristics and the uninsured rate changed between 2013 and 2014, so the aggregate model cannot be extended past 2013.

The second approach uses a linear probability difference-in-difference regression, hereafter referred to as the person-level model, to control for other factors that might have affected the uninsured rate in addition to the questionnaire change (Equation 2). The person-level model also enables us to examine differences in the predicted uninsured rate between subgroups, something that we could not do using the aggregate-level approach due to sample size. This person-level model assumes that certain effects are constant over time, but we also test this assumption using interactions between key demographic characteristics and a time component. In Subsection 6.2., Appendix 2, the variance formulas and the weighting procedure are discussed for the person-model.

$$P(UNINS_i) = \alpha + \gamma_t + \beta_1 CPS + \beta_2 \text{CPS} * \text{QCHANGE}_{CPS} + \beta_3 X + \varepsilon \qquad (2)$$

Where UNINS is a dummy for being uninsured (i.e., the probability of being uninsured is the average between the ACS and the CPS), γ_t is calendar year (controlling for the effect of the year), CPS is an indicator for the data used (CPS = 1 if CPS data are used and CPS = 0 if ACS data are used), $\text{QCHANGE}_{\text{CPS}}$ is a dummy representing the redesigned survey questions (for CPS in 2013 and beyond), X is a vector of covariates (listed below) that are controlled for in the model, and i denotes the individual. In this application, the primary parameter of interest is β_2; it represents the effect of the questionnaire change in the CPS. The regression is estimated separately for the full sample and by subgroup. At the individual level, we use the estimated coefficients to predict the probability of being uninsured for each individual in the data set. Now, we have a predicted individual value for the traditional questions by subtracting the parameter that is due to the questionnaire change (β_2) (see Equation 3). By subtracting the effect of the questionnaire effect from the redesigned estimate, this model also produces a predicted estimate of the traditional uninsured rate of the CPS in 2009 through 2014. All ACS data are removed from the analysis at this point.

$$\textit{if year} \geq 2013 \textit{ then } Pred(\textit{Traditional Question}_i) = Pred(UNINS_i) - \beta_2$$

$$\textit{else } Pred(\textit{Traditional Question}_i) = Pred(UNINS_i) \qquad (3)$$

After the individual prediction of the traditional questions is done, we take the weighted mean for our final national estimate, by year. In the regression, we control for age, race, sex, Hispanic origin, disability status, citizenship, receipt of Supplemental Nutrition Assistance Program (SNAP) benefits, income-to-poverty ratios, and living in a metropolitan statistical area. In addition to the above characteristics, we control for several interactions between terms. Details are listed in Subsection 6.1., Appendix 1.

We denote this collection of control variables as X in Equation 2. Without these covariates in the model, the results of the person-level model would be equivalent to those of the aggregate-level approach (Equation 1) if the person-level model had only used

2009–2013 data (the model uses 2009 to 2014 data). We use the aggregate-level model as a robustness check for the person-level model, since the simplest form of the person-level model is very close to the aggregate-level model. If the person-model was based on just 2009–2013 data and $B_3 = 0$ (i.e., no covariates except for time, survey, and the years that the CPS redesigned survey questions were in effect), then the person-level estimate of the uninsured in 2013 should be very close to the aggregate model. This robustness check was done and the intuition is borne out in the data (not shown).

4. Results

The first part of the analysis uses the aggregate method to estimate what the uninsured rate would have been, using the traditional questions (i.e., had the questions not changed). The time trend in the estimated uninsured rate for each survey is shown in Figure 1. Recall that one of the primary reasons the CPS was redesigned was to address the overestimate of the uninsured rate compared to other federal surveys. The uninsured rate in the traditional CPS was higher than in the ACS, despite the fact that the uninsured in the CPS rate reflects the entire calendar year, while the uninsured rate in the ACS reflects a point in time. The uninsured rate in the redesigned CPS, on the other hand, is lower than the uninsured rate in the ACS, which is consistent with the expectations that someone is less likely to be uninsured for an entire calendar year than on any given day in a year.

The aggregate model predicted the uninsured rate in 2013 to be 15.3% (Table 2). The predicted year-to-year change in the uninsured rate from 2012 to 2013 is, by design, equivalent to the change in the uninsured rate in the ACS over that period: 0.1 percentage point (not statistically significant). As the 2013 CPS uninsured rate using the redesigned questions was 13.3%, the estimated questionnaire effect was 1.9 percentage points.

The next part of the analysis uses person-level models to predict the questionnaire effect and the traditional uninsured rate in 2013, while controlling for possible confounding covariates. Because the fully specified model has 90 variables and there are 21 subgroups, we do not show all of the regression parameters (available upon request). The person-level model shows that the estimate of the questionnaire effect was 1.7 percentage points (15.0% for the predicted traditional CPS less 13.3% for the redesigned CPS), not statistically different from the questionnaire effect derived from the aggregate-level model (see Table 2). The predicted uninsured rate for 2013 using the person-level model is 15.0%, slightly lower than the person-level prediction in 2012. In addition, when comparing the predicted estimates for 2013, we find that the person-level model produces a slightly lower estimate than the aggregate-level model (see Subsection 6.2., Appendix 2 for a discussion of how the standard errors were calculated). Overall, the predictive power of the person-level model was 21%. The degree of the predictive power varied by subgroup. For example, the population aged 65 and over had an R-squared of 2%, which was in line with the expectation that this group would be relatively unaffected by the questionnaire change.

The person-level model allows us to examine variation in both the questionnaire effect (in 2013) and change in the uninsured rate (2012 to 2013) by subgroup, including race and Hispanic origin, age, low-income status, and labor force status (for adults aged 19 to 64). The data show variation in the effect of the questionnaire change by subpopulation group,

Table 2. Observed and predicted uninsured rates, by year and survey.

	Observed						Predicted traditional CPS			
	Traditional CPS		Redesigned CPS		ACS		Aggregate model		Person-level model	
Year	Percent	SE	Percent	SE	Percent	SE	Percent	SE	Percent	SE
2009	16.1	(0.14)			15.1	(0.05)			16.2	(0.08)
2010	16.3	(0.15)			15.5	(0.05)			16.2	(0.08)
2011	15.7	(0.12)			15.1	(0.04)			15.7	(0.08)
2012	15.4	(0.13)			14.8	(0.04)			15.4	(0.08)
2013			13.3	(0.12)	14.5	(0.04)	15.3	(0.08)	15.0	(0.08)
2014			10.4	(0.13)	11.7	(0.04)			12.3	(0.08)

Source: 2009–2014 one-year American Community Surveys (restricted data), 2010–2015 Current Population Survey Annual Social and Economic Supplements.

Note: Traditional CPS refers to the official uninsured estimate from the CPS for calendar years from 2012 and before. Redesigned CPS refers to the official uninsured estimate from the CPS for calendar years 2013 and beyond.

Note: The effect of questionnaire change in 2013 is calculated as follows. For the aggregate model, the questionnaire effect is 1.9 percentage points and is calculated by subtracting the redesigned CPS estimate (13.3) from the predicted traditional CPS estimate (15.3%). For the person-level model, the questionnaire effect is 1.7 percentage points, which is β_2 in Equation. 2.

Table 3. Predicted questionnaire effect and year-to-year change for subgroups, person-level model, 2012 to 2013.

	Questionnaire effect			Predicted uninsured rate						
	2013			2012		2013		Diff. 2012 to 2013		
	β_2	SE		%	SE	%	SE	% pts	SE	
Total Population	−1.7	−0.04	***	15.4	−0.08	15.0	−0.08	0.4	−0.12	**
Race and Hispanic origin										
Non-Hispanic White	−1.0	−0.05	***	10.9	−0.10	10.6	−0.10	0.3	−0.15	+
Non-Hispanic Black	−3.0	−0.21	***	18.7	−0.29	18.5	−0.32	0.2	−0.43	
Hispanic	−3.6	−0.16	***	29.4	−0.22	28.1	−0.22	1.2	−0.31	***
Age										
Under 65	−2.0	−0.05	***	17.6	−0.09	17.2	−0.09	0.4	−0.13	**
Under 19	−1.5	−0.09	***	9.2	−0.16	8.9	−0.15	0.3	−0.22	
Aged 19 to 64	−2.2	−0.06	***	21.0	−0.12	20.6	−0.13	0.4	−0.18	*
Aged 65 and over	−0.2	−0.04	***	1.5	−0.19	1.6	−0.17	−0.1	−0.25	
Low income by age										
All ages	−3.8	−0.12	***	26.5	−0.18	25.9	−0.18	0.7	−0.25	**
Under 65	−4.3	−0.15	***	30.3	−0.20	29.5	−0.20	0.8	−0.28	**
Under 19	−2.7	−0.17	***	13.6	−0.26	13.1	−0.26	0.5	−0.37	
Aged 19 to 64	−5.2	−0.20	***	40.2	−0.30	38.9	−0.31	1.3	−0.43	**
Aged 65 and over	−0.8	−0.10	***	2.7	−0.37	2.9	−0.35	−0.1	−0.51	
Labor force status (19–64 year olds)										
Full time	−1.8	−0.07	***	16.9	−0.17	16.7	−0.19	0.2	−0.25	
Part time	−2.4	−0.23	***	24.2	−0.44	23.8	−0.41	0.4	−0.60	
Unemployed	−3.2	−0.48	***	40.5	−0.88	38.1	−0.94	2.4	−1.29	+
Not in labor force	−3.3	−0.18	***	23.3	−0.34	23.2	−0.3	0.1	−0.48	

Source: Authors' calculations from 2013–2014 Current Population Survey Annual Social and Economic Supplements.
Note: ***p < .001, **p < .01, *p < .05, + p < .1 for t-test.

but little variation in the year-to-year change in the uninsured rate (Table 3). Among the race and Hispanic origin subgroups, the effect of the questionnaire change on the uninsured rate was the lowest for non-Hispanic Whites, who experienced a change of almost half of the effect on the total population. Non-Hispanic Blacks and Hispanics, on the other hand, had decreases in their uninsured rates due to the questionnaire change that were greater than the average for the population. Non-Hispanic Whites and Hispanics experienced increases in their uninsured rates from 2012 to 2013, though the change was greater for Hispanics.

The age group with the largest questionnaire effect on the uninsured rate was adults aged 19 to 64. Children under age 19 had a difference that was about half of the change for the working-age adults, and adults aged 65 and over showed the smallest difference in the uninsured rate between the traditional and redesigned questions. When examining the low-income population (family income is less than or equal to 200% of the Income-to-Poverty-Ratio), the effect of the questionnaire change is about twice as large as the effect for the total population. The age pattern of the questionnaire effect on the uninsured rate for the low-income population is consistent with the age pattern for the total population: the smallest effects are associated with the elderly and the largest effects are associated with working-age adults. In these low-income groups, only the low-income population overall, working-age adults, and people under 65 experienced a change in the uninsured rate from 2012 to 2013.

We also examined labor force status for working-age adults and found that the uninsured rate differed less for workers than nonworkers when the questionnaire changed. The questionnaire change had the largest effect for the unemployed and adults not in the labor force, a smaller effect for part-time workers, and the smallest effect on the uninsured rate for full-time workers. Between 2012 and 2013, unemployed adults were the only labor force category to experience a change in the uninsured rate.

5. Conclusion

In 2014, new health insurance questions replaced the traditional questions in the CPS. This change established a disruption in the time trend for the CPS, where estimates from 2012 and earlier cannot simply be compared to the estimates from 2013 and beyond without disaggregating the effect of the questionnaire change from the time trend. While the redesigned questions improved health insurance estimates by making them more in line with other federal surveys, it is also important to maintain the historical time trend, so that researchers and policymakers can take a broader view of trends in health insurance coverage from before the ACA's implementation in 2010. This article fills that gap, by predicting what the uninsured rate would have been if there was not a change in the health insurance questions.

Using the year-to-year change in the uninsured rate as measured by the ACS, together with the uninsured rate from the traditional questions in the CPS in 2009 through 2012, we predicted that the uninsured rate would have been 15.3% in 2013 using aggregate-level data, unchanged from the level in 2012. Using person-level data, we employed a difference-in-difference model to control for demographic and socioeconomic changes in the population, and predicted that the uninsured rate would have been 15.0% in 2013, not

different from the aggregate-level prediction but slightly lower than the person-level prediction in 2012. Both the aggregate- and person- level models could be used to evaluate other survey redesigns. For the former application, additional research would be needed to determine just how far the model could be extended without over-specifying the model. For the latter, one would need to update the data every couple of years and evaluate the goodness of fit to assure that the bridge still fit.

In addition to continuing the historic time trend, we also estimated what effect changing the CPS questionnaire had on the uninsured rate in 2013, the first year of data that collected health insurance using the redesigned questions. We found that the redesigned questions reduced the uninsured rate by about 1.7 percentage points using the person-level model to control for confounding covariates (not statistically different from the aggregate-level estimate). The questionnaire effect varied by subpopulation, and was greater, in general, for the groups that had higher uninsured rates in 2012 as measured by the traditional questions. For instance, the questionnaire change had the smallest effect on the uninsured rate for children, seniors, and adults working full time. These findings were consistent with the expectation that populations with higher rates of coverage would be less affected by the questionnaire change. Another finding was that the effect of the questionnaire change for non-Hispanic Whites was lower than average and lower than for the other race and Hispanic origin groups. This finding may be explained by a disproportionate number of non-Hispanic Whites over age 65, which is consistent with previous research (Day 2013). Finally, we found that the questionnaire change had the largest impact for low-income working-age adults. This means that low-income adults are reporting insurance more often using the redesigned questions as compared with the traditional questions.

Due to changes in the demographic and socioeconomic composition of the population between 2012 and 2013, the aggregate-level model produced a different predicted uninsured rate in 2013 than did the person-level model. These changes in the population also explain why the aggregate-level model did not produce a change in the uninsured rate between 2012 and 2013, while the person-level model showed a slight decrease. Therefore, when possible, it is preferable to use a difference-in-difference regression on person-level data to control for overall population changes, but when it is not possible, aggregate-level data produce similar results.

It is important to note that the person-level model is a linear probability model instead of a logit model. We chose to use a linear probability model because it is the standard model in the difference-in-difference context. However, we could have used a logit model.

In this article, we take advantage of the stable relationship between estimates from multiple surveys over time in order to fill in the gap during a disruption in the time series.

Surveys need to change questions for many reasons: to reduce respondent burden, improve validity, and to harmonize questions with other surveys. When survey questions change, there is always a balance between maintaining the time trend and improving the questions. However, as long as there are other sources of data that track in parallel over time with the survey, there does not have to be a tradeoff. This approach can be used when other surveys change questions.

6. Appendix

6.1. Appendix 1: Covariates Included in Person-Level Model

Table A1. Covariates included in person-level model.

Demographic covariates	Health covariates	Interaction terms
Age	Disability status	IPR by survey[1]
0–5	Has a disability	Age 19–25 by year[2]
6–18	Does not have a disability	Receives SNAP by 0–138% IPR[3]
19–25		IPR by year[4]
26–34	**Socioeconomic covariates**	SNAP by year[4]
35–44	Supplemental Nutrition	SNAP by 0–138% IPR by year[4]
	Assistance Program (SNAP)	
45–54	Receives SNAP benefits	
55–64	Does not receive SNAP benefits	
65–74	Income-to-poverty ratio (IPR)	
75 and over	0–138% IPR	
Sex	139–199% IPR	
Male	200–299% IPR	
Female	300–399% IPR	
Race and Hispanic Origin	400–499% IPR	
Non-Hispanic White alone	500% and over IPR	
Non-Hispanic Black alone		
Non-Hispanic other Hispanic	**Geographic covariates**	
	Metropolitan Statistical Area (MSA)	
Citizenship	Lives in MSA	
Citizen	Does not live in MSA	
Not a citizen	Medicaid Expansion	
	State (as of 1/1/14)	
	Lives in expansion state	
	Does not live in expansion state	

Notes:

[1]We include interactions between each IPR level and the survey because income is measured differently between the two surveys (the CPS collects over 50 types of income while the ACS collects only eight).

[2]Interactions between the age group 19 to 25 and each calendar year account for changes in the relationship between health insurance coverage and age over the period due to the 2010 implementation of the dependent coverage provision of the ACA.

[3]An interaction between receipt of SNAP benefits and IPR less than or equal to 138% of the poverty threshold is included because only low-income families are eligible for SNAP benefits.

[4]We control for time-effects (in addition to the direct effect of γ_t), by including interactions for year by all levels of the IPR, year by receipt of SNAP benefits, and a three-way interaction between year, receipt of SNAP benefits, and low-income status (IPR is between 0 and 138% of the poverty threshold).

6.2. Appendix 2: Variance Estimation

All of the estimates presented in this article, including both the mean and the variance, are weighted estimates. In the article, we focus on the means of the uninsured. Appendix 2 focuses on the modeled variance. We will not focus on direct estimates of variance. Direct estimates of the means and variances use the person data and the post-stratified weights and the standard errors are computed using successive difference replication (Fay and Train 1995).

Aggregate Model, Tabular

For the aggregate model, the Rate and Var(Rate) is calculated directly from the survey. The second step is estimating the variance of the standard error of the prediction.

$$SE_{Pred(CPS2013)} = \text{SQRT}\left(Var(Rate_{ACS2013}) + \frac{1}{16} \sum_{k=2009}^{2012} (Var(Rate_{ACS}) + Var(Rate_{CPS}))_k \right) \quad \text{(A1)}$$

Person-Model, Regression

The sample size for the ACS is roughly 35 times the sample size of CPS. If we ignore the sample size difference, the contribution of the ACS to the final prediction in the CPS estimate will be overstated. Therefore, we must adjust the final estimates of the mean and the variance. We do this in steps: first, we calculate a person-weight adjustment, and second, we use the adjusted person weight to calculate the variance of the prediction.

Step 1: Person-Weight Adjustment

To account for the complex survey designs in both the ACS and CPS in the regression-based person-model, each year of data for each survey is weight-adjusted to match their effective sample size (Kish 1965) for the uninsured rate. The Rate and Var(Rate) is calculated directly from the survey. The ESS is calculated (Equation A2a) on the uninsured rate.

$$ESS_j = (Rate_j)(1 - Rate_j)/Var(Rate_j) \quad \text{(A2a)}$$

where j = calendar year, subgroup, and survey.

An adjustment factor is applied (Equation A2b).

$$Adjust_j = ESS_j / Sum\ of\ person\ weights_j \quad \text{(A2b)}$$

An adjusted/final person-weight (FPW) is used for the regression (Equation A2c).

$$Final\ Person\ Weight_{i,j} = Adjust_j \times Person\ Weight_i \quad \text{(A2c)}$$

Where i = person in group j.

Step 2: Final Estimate of Variance

Using the FPW as the weight for the regression model, we run the model. For each observation that comes from the CPS, the relevant output for calculating the standard error of the traditional questions ($SE(\hat{Y}_j)$) are: FPW, prediction error under the model ($SE(\hat{Y}_{i,j})$), calendar year, and subgroup. All of the ACS observations are dropped. At the person-level, calculate an adjustment to the person-level variance of the prediction of being uninsured.

$$P1\_VAR_{i,j} = FPW_{i,j}^2 \times SE(\hat{Y}_{i,j})2 \tag{A3}$$

Equation A4 is the standard error of the weighted predicted mean for our final national estimate, by year and subgroup.

$$SE(\hat{Y}_j) = \text{SQRT}\left(\sum P1\_VAR_{i,j}\right) / \sum FPW_{i,j}) \tag{A4}$$

7. References

Bailey, J. and A. Chorniy. 2016. "Employer-Provided Health Insurance and Job Mobility: Did the Affordable Care Act Reduce Job Lock?" *Contemporary Economic Policy* 34(1): 173–183. Doi: http://dx.doi.org/10.1111/coep.12119.

Blewett, L.A. and M.E. Davern. 2006. "Meeting the Need for State-Level Estimates of Health Insurance Coverage: Use of State and Federal Survey Data." *Health Services Research* 41(3): 946–975. Doi: http://dx.doi.org/10.1111/j.1475-6773.2006.00543.x.

Bradley, C.J., D. Neumark, and M. Motika. 2012. "The effects of health shocks on employment and health insurance: the role of employer-provided health insurance." *International Journal of Health Care Finance and Economics* 253–267. Doi: http://dx.doi.org/10.1007/s10754-012-9113-2.

Call, K.T., G. Davidson, M. Davern, E.R. Brown, J. Kincheloe, and J.G. Nelson. 2008. "Accuracy in self-reported health insurance coverage among Medicaid enrollees." *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* 45(4): 438–456. Doi: http://dx.doi.org/10.5034/inquiryjrnl_45.04.438.

Centers for Medicare and Medicaid Services. 2014. *National Health Expenditure Accounts: Methodology Paper, 2013*. Centers for Medicare and Medicaid Services. Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/dsm-13.pdf (accessed March 2018).

Davern, M., H. Rodin, L.A. Blewett, and K.T. Call. 2007. "Are the Current Population Survey Uninsurance Estimates Too High? An Examination of the Imputation Process." *Health Services Research* 42(5): 2038–2055. Doi: http://dx.doi.org/10.1111/j.1475-6773.2007.00703.x.

Day, J. 2013. "Medicare and Medicaid, Age and Income." *Random Samplings*. U.S. Census Bureau. Available at: https://www.census.gov/newsroom/blogs/random-samplings/2013/09/medicare-and-medicaid-age-and-income-2.html (accessed March 2018).

Fay, R.E. and G.F. Train. 1995. "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties."

Proceedings of the American Statistical Association Conference. Orlando, August 13–17, 1995. Available at: https://www.census.gov/content/ dam/Census/library/working-papers/1995/demo/faytrain95.pdf (Accessed March 2018).

Kaestner, R. and D. Lubotsky. 2016. "Health Insurance and Income Inequality." *The Journal of Economic Perspectives* 30(2): 53–77. Doi: http://dx.doi.org/10.1257/jep.30.2.53.

Kenney, G. and V. Lynch. 2010. "Monitoring Children's Health Insurance Coverage Under CHIPRA Using Federal Surveys." Chapter. 8, In *Databases for Estimating Health Insurance Coverage for Children: A Workshop Summary*, edited by Thomas Plewes, 65–82. Washington, DC: National Academies Press. Doi: http://dx.doi.org/10.17226/13024.

Kish, L. 1965. *Survey Sampling*. New York: John Wiey & Sons, Inc.

McGeary, K.A. 2009. "How Do Health Shocks Influence Retirement Decisions?" *Review of Economics of the Household* 7(3): 307–321. Doi: http://dx.doi.org/10.1007/s11150-009-9053-x.

O'Hara, B. and K. Caswell. 2013. *Health Status, Health Insurance, and Medical Services Utilization: 2010*. Current Population Reports, P70-133RV. Washington: U.S. Census Bureau.

Pascale, J. 2016. "Modernizing a Major Federal Government Survey: A Review of the Redesign of the Current Population Survey Health Insurance Questions." *Journal of Official Statistics* 32(2): 461–486. Doi: http://dx.doi.org/10.1515/jos-2016-0024.

Pascale, J., M.I. Roemer, and D. Resnick. 2009. "Medicaid Underreporting in the CPS: Results from a Record Check Study." *Public Opinion Quarterly* 73(3): 497–520. Doi: http://dx.doi.org/10.1093/poq/nfp028.

Pascale, J., M. Boudreaux, and R. King. 2016. "Understanding the New Current Population Survey Health Insurance Questions." *Health Services Research* 51(1): 240–261. Doi: http://dx.doi.org/10.1111/1475-6773.12312.

Smith, J.C. and C. Medalia. 2014. *Health Insurance Coverage in the United States: 2013*. Current Population Reports, P60-250, U.S. Census Bureau. Washington, DC: U.S. Government Printing Office.

Smith, J.C. and C. Medalia. 2015. *Health Insurance Coverage in the United States: 2014*, Current Population Reports, P60-253, U.S. Census Bureau, Washington, DC: U.S. Government Printing Office.

Received January 2017
Revised December 2017
Accepted April 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 203–229, http://dx.doi.org/10.2478/JOS-2019-0010

Adjusting for Measurement Error in Retrospectively Reported Work Histories: An Analysis Using Swedish Register Data

Jose Pina-Sánchez[1], *Johan Koskinen*[2], *and Ian Plewis*[2]

We use work histories retrospectively reported and matched to register data from the Swedish unemployment office to assess: 1) the prevalence of measurement error in reported spells of unemployment; 2) the impact of using such spells as the response variable of an exponential model; and 3) strategies for the adjustment of the measurement error. Due to the omission or misclassification of spells in work histories we cannot carry out typical adjustments for memory failures based on multiplicative models. Instead we suggest an adjustment method based on a mixture Bayesian model capable of differentiating between misdated spells and those for which the observed and true durations are unrelated. This adjustment is applied in two manners, one assuming access to a validation subsample and another relying on a strong prior for the mixture mechanism. Both solutions demonstrate a substantial reduction in the vast biases observed in the regression coefficients of the exponential model when survey data is used.

Key words: Bayesian statistics; measurement error; mixture model; retrospective data; unemployment.

1. Introduction

Many different forms of measurement error (ME) have been treated in the literature (Berkson 1950; Black et al. 2000; Neuhaus 1999; Novick 1966). Moreover, in some instances, different forms of ME interact within the same data-recording strategy. We consider here such a case, motivated by a retrospective data set, where ME is best modelled using a combination of errors. We demonstrate the application of a mixture of ME using validation samples of 20% and 40%, as well as in the absence of validation samples.

Retrospective questions are useful for collecting information about life-course events over a span of time from a single contact with a respondent. They are cheaper to administer than alternative data collection schemes that rely on repeated updates of the current state (Solga 2001). On the other hand, retrospective questions are prone to ME due to memory failures. These memory failures can take different forms.

For questions retrieving specific events, for example, "*When was the last time you went to the dentist?*" (Office for National Statistics 2006) – we can detect *telescoping effects*

[1] School of Law, University of Leeds, Leeds, LS2 9JT, United Kingdom. Email: j.pinasanchez@leeds.ac.uk
[2] Social Statistics, University of Manchester, Manchester, M13 9PL, United Kingdom. Emails: Johan.Koskinen@manchester.ac.uk and Ian.Plewis@manchester.ac.uk
Acknowledgments: We want to thank Sten-Åke Stenberg for granting us access to the "Longitudinal Study of the Unemployed", a unique data set without which this research project would not have been possible.

(Golub et al. 2000; Johnson and Schultz 2005). Neter and Waksberg (1964) coined this term to refer to the temporal displacement of an event, whereby people perceive recent events as being more remote than they are (backward telescoping or time expansion) and distant events as being more recent than they are (forward telescoping or time compression). Other researchers (Bradburn et al. 1994; Huttenlocher et al. 1988; Rubin and Baddeley 1989) have argued that, rather than distorted time perceptions, recall errors take the form of random ME around the reported date with the size of the error being proportional to the distance between the time of the interview and the actual date. That is, the further away the date of the event to be reported, the harder it is to recall and therefore the bigger the ME.

For questions retrieving count data – that is, those enquiring about the number of times an event has been experienced over a period of time, for example, *"How many times during the last two years have you put together self-assembly furniture at home?"* (Office for National Statistics 2008). – *interference effects* have been detected. This term was coined by Crowder (1976) and refers to the probability of recalling a particular event being inversely related to the number of times the respondent experiences similar events (Shiffrin and Cook 1978).

All these types of memory failure result in a significant loss of reliability and validity in most of the variables collected retrospectively, which in turn can have severe consequences in the form of loss of statistical power and biased estimates when used in statistical models. Nonetheless, although undoubtedly inconvenient, these effects can be mitigated through the implementation of methods for the adjustment of ME. In particular, much of the literature has focused on the implementation of adjustments where multiplicative models (see Section 2) are used to specify the distribution of recall errors (Augustin 1999; Dumangane 2007; Glewwe 2007; Holt et al. 2011; Pickles et al. 1996, 1998; Pina-Sánchez 2016; Skinner and Humphreys 1999).

Multiplicative models can be successfully used to map different ME processes stemming from memory failures in the reports of specific events. However, things become much more complicated when dealing with retrospective reports of different kinds of histories. Such reports require the identification and timing of the same or different events in chronological order. Take this question from the 2003 Improving Survey Measurement of Income and Employment project (Jenkins and Lynn 2005) as an example, *"Have you received Job Seeker's Allowance at any time since <DATE OF PREVIOUS INTERVIEW>?"* If yes, interviewees were then asked *"For which months since <MONTH OF INTERVIEW> have you received Job Seeker's Allowance?"*. Here, we might expect spells of receiving the benefit having misdated start or end times. However, we should also expect ME in the form omitting spells that did occur, over-reporting spells that did not occur and misclassifying spells by, for example, reporting receipt of one kind of benefit when it was, in fact, a different benefit. These other types of errors can generate severe distortions since, unlike misdated starts or ends of spells, they give rise to durations of spells that are unrelated to their true values. Under such circumstances, the impact of ME on the estimates of, say, a regression model, should be expected to be more serious than when using spells solely subject to multiplicative errors, while the potential application of adjustment methods is more limited and more complex to implement.

In this article, we use administrative data from the Swedish register of unemployment linked to survey reports of individuals' work histories (described in Section 3). Under the assumption that the former is perfectly measured, we assess: a) the prevalence and forms of ME in the retrospective reports of work histories (Subsection 4.1), b) the consequences of using durations of spells of unemployment captured from those work histories as the response variable in an exponential model (Subsection 4.2), and c) the effectiveness of different approaches based on Bayesian methods to adjust for the consequences of ME in such an exponential model (Section 5). In particular, we demonstrate the potential of an adjustment using a Bayesian mixture model. We calibrate the ME model separately using validation samples and by relying on past information. Section 6 concludes with a list of recommendations regarding the collection of work histories retrospectively, and with a discussion of possible further improvements in the adjustment model, using auxiliary data.

2. Modelling Recall Errors

In order to adjust for the consequences of using variables prone to ME, many analysts rely on the assumption that the error mechanism is classical. The classical ME model was first formally defined by Novick (1966) as $X^* = X + V$; where X^* is the observed variable, equal to the true variable, X, plus the ME term, V, with the following five assumptions:

Table 1. Assumptions of the classical model.

$E(V) = 0$;	null expectancy
$Var(V_i) = Var(V)$;	homoscedasticity
$V \sim N\,(0,\ Var(V))$;	normally distributed
$Cov(X,\ V) = 0$;	indep. error and true value
$E(Y\,\vert X,\ X^*) = E(Y\,\vert X)$;	non-differentiality

The classical model nicely reflects the type of ME that we can expect to find in continuous variables prone to random errors, for example when measuring temperature using an unreliable thermometer. In addition, the classical model is often used as the foundation upon which more complex ME processes are specified. One such a case is the previously mentioned multiplicative model, where the additive relationship between the true value and the error is substituted by a multiplicative one, so, $X^* = XV$.

The same assumptions about the error term described above apply, except that the error term is now log-normally distributed with a mean of one. The ME has a symmetric effect across the true values and maintains the scale used in duration and count data, from 0 to ∞. More importantly, since the effect of V on X^* is proportional to the value of X, the multiplicative model can be used to reflect a random ME process stemming from memory failures observed in retrospective questions capturing events or counts of events (see the examples from the Office for National Statistics in the introduction). That is, the greater the number of events experienced, or the further the event from the time of the interview, the higher the prevalence of ME. Note, as well, that this same model can also be used to account for systematic (i.e., nonrandom) recall errors, such as those found under the presence of backward and forward telescoping effects, by shifting the distribution of V to the right or left so its mean goes below or above one.

However, this model cannot adequately account for the type of ME observed in questions where interviewees are requested to report retrospectively not only the duration of a specific spell, but an entire history. As anticipated in the introduction, such reports are subject to additional forms of ME. In particular, the spells comprising the reported history will not only be affected by problems of their start or end being misdated, or any of the other typical forms of bias found in survey data, such as social desirability, or acquiescence bias, but also by more problematic issues of omission/overreporting and misclassification. Each of these forms of ME can distort the true durations of the spells in different ways. In Figure 1, we present three examples of the potential implications of ME when the retrieved durations are considered as the response variable of an event history model based on first spells (as opposed to multiple spells). The examples are taken from the data set presented in the following section, where all subjects start from a state of unemployment and the window of observation covers 395 days. The true spells of unemployment are now denoted by Y and are represented by the continuous lines. The error is represented by V and is encompassed by the bracket immediately below, whereas the observed durations are denoted by $Y*$.

When spells are misdated, the observed durations can appear shorter or longer than the true ones. This is represented by the first case shown in Figure 1, where the only spell of unemployment experienced within the window of observation has been reported to be 90 days longer than it really was. Misdating errors are the only types of recall errors that can occur when subjects known to be in one particular state are requested just to report the duration or the end date of that spell. This simple case could be well represented by the classical multiplicative model, where shorter durations will be associated with more accurate reports.

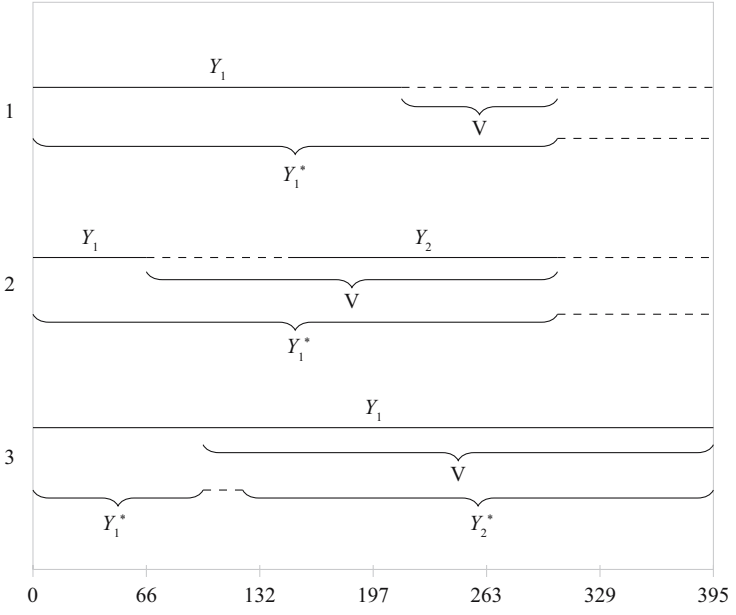

Fig. 1. Work history durations affected by different types of measurement error.

The omission of spells could more seriously distort estimation based on this data. Take the second work history presented in Figure 1, for example: if the spell representing a status different from unemployment starting in day 66 was omitted, the two spells of unemployment that occur before and after would be artificially linked, and the reported work history for that subject would look like a unique spell of unemployment. The same situation could also be reproduced as a result of misclassification of spells, specifically if the spell between Y_1 and Y_2 was a false positive case of unemployment (e.g., a registered case of employment reported as a spell of unemployment). An equally problematic form of ME occurs as a result of overreported spells. The third case in Figure 1 represents an example of a spell different from unemployment being mistakenly reported. This spell only covers 20 days (from day 100 to 120), but it has a severe effect, since it splits a true duration that was right censored and generates an observed duration only 100 days long.

Finally, we could also categorise a ME effect from misclassified spells in the form of false negative spells of unemployment. Given our particular setting, where all work histories start from unemployment and only first spells are considered, the effect of false negative spells will resemble that of missing data since these cases will be completely unobserved. The difference lies in the fact that in the presence of missing data we know which cases are missing. As long as the probability of committing false negatives is independent of the duration of unemployment and other observed explanatory variables, it will only affect the precision of model estimates.

3. Data

The data we use has been obtained from the "Longitudinal Study of the Unemployed" a research project designed by the Swedish Institute for Social Research (SOFI) at Stockholm University, directed by Sten-Åke Stenberg, and with the collaboration of the register of unemployment (PRESO). This register provided individual-level data on the work status of the participants of three surveys, run in 1992, 1993, and 2001. The sample was designed to capture 830 jobseekers randomly selected amongst those registered as unemployed on 1992-02-28 who met the following criteria: aged between 25 to 55 years, of Nordic nationality, no occupational disability, and seeking a full-time job. The three surveys are relatively similar with respect to the composition of both the sample of participants and the questionnaire, although for reasons of attrition the response rate for the three surveys decreased across time from 64.7%, to 59.4% and 50%. In this study, we use data derived from a retrospective question on work status from the 1993 survey, which reads as follows:

Which of the alternative answers best describes your main activity the first week of 1992? When did this activity start? When did it end?

Which was the subsequent main activity? When did this activity start? When did it end?

In order to simplify the reported work histories to be analysed, we set the beginning of the window of observation at 1992-02-28 and only consider subjects who started from a state of unemployment in both the register and the survey. Under this restriction, our sample mimics the structure seen in state-based samples (Holt et al. 2011), where the sample frame is created out of individuals who are known to be in a particular state. Our final sample size captures 381 individuals (from a total of 532 captured by both survey and

Table 2. Descriptive statistics of the sample.

	Mean/ Median	Standard deviation/ Interquartile range	Minimum	Maximum
Age	37	8.8	26	55
Experience	2.6	0.6	1	3
Register durations*	253	303	1	395
Survey durations*	92	144	1	395

*Since these variables are subject to censoring, their medians and interquartile ranges are reported.

register). The window of observation encompasses all spells from 1992-02-28 to 1993-03-30, where the end date represents the earliest day that interviews were held for the second wave of the survey. Right censoring is present in both the survey and register data sets.

In addition to drawing the duration of spells of unemployment from PRESO, the register also provides the age and experience of the 381 subjects. Given that *age* is an important variable in the register, the probability that it is prone to ME is very low. This is not so for *experience*, which captures self-reported levels of experience in the type of work that the subject applied for on a scale with three levels (low, medium, and high). However, in our analysis we assume that both *age* and *experience* are free of ME. The value for *age* is taken in February 1993, which gives us a mean sample age of 37 and a standard deviation of 8.8, while for *experience* the mean of the monthly reported levels in 1992 is used, with a mean and a standard deviation in our sample of 2.6 and 0.6 respectively. The rest of the descriptive statistics for the variables used in our study are reported in Table 2.

4. Prevalence and Impact

In a study using this same data set, Pina-Sánchez et al. (2014) found evidence of the different types of errors discussed in Section 2. For example, 57% of the first spells reported were misdated by more than 31 days, and 30% were misclassified (interviewees reported to be employed or out of the labour force when, in fact, they were registered as unemployed, or vice versa). In addition, a tendency to omit spells of unemployment was detected since the ratio for the mean number of spells of unemployment reported over those registered during the window of observations was 1.4/1.7.

Here, we carry forward this analysis to assess the prevalence and effect of the ME found in spells of unemployment retrieved from work history reports. First, we look at differences between spells of unemployment from the same subjects and across the same window of observation captured by the survey and the register to estimate the prevalence of ME in the former. Second, we assess the impact of such ME on event history analysis, by looking at differences in regression coefficients and their measures of uncertainty between the same models when survey instead of register data is used. These analyses are based on the key assumption that spells from the register are free of ME. This is a realistic, yet not entirely perfect, assumption. Even data from registers can be affected by ME (Kapteyn and Ypma 2007; Pavlopoulos and Vermunt 2015). As a result, in the few

instances where the register is inaccurate, we might be wrongly considering spells of unemployment to be misreported.

4.1. Prevalence of Measurement Error in Durations of Unemployment

A comparison of medians (Table 2) shows that the durations of spells of unemployment are substantially longer in the register (253 days) than in the survey (92 days). This longer duration of registered spells is also reflected by the 133 cases (35% of the sample) that are right censored (remained unemployed by the end of the window of observation), compared with only 23 (6% of the sample) in the survey. These differences can be appreciated in Figure 2(a), where the survivor functions for the two types of durations are represented using Kaplan-Meier estimates (P(S) indicates the probability of not having made a transition out of unemployment at a particular time). The two data sets show a similar path for the first 30 days; from that point until about day 100 the two measures diverge due to an accelerated failure rate in the survey; from then on, the two survivor functions behave roughly similarly and the gap between them is maintained.

The different failure rates observed for the reported and registered durations indicate the presence of a systematic component in the ME process. In Figure 2(b) we plot the density of the error term assuming it is multiplicative, $V = Y^*/Y$. Here we can observe that, although a majority of errors are centred around one (as could be expected in a classical multiplicative framework), the distribution is bimodal and shows a substantial number of extreme values. The calculation of the error term might be biased since the right-censored cases were taken to be equal to 395. However, this could not account for the extreme values seen here. Hence, we can conclude that the ME process is not multiplicative.

The ME can also be assessed using scatter (c) and density plots (d). The former can be used to assess the effect of ME on a case-by-case basis, while the latter represents the

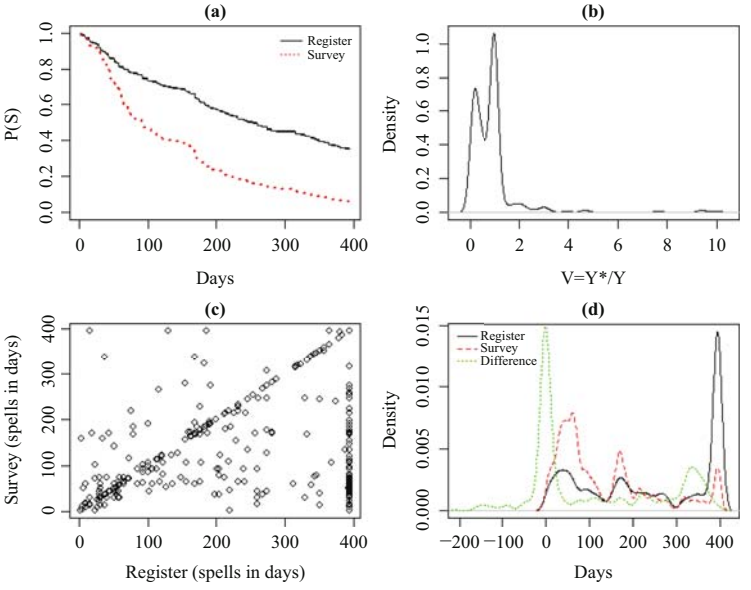

Fig. 2. Survey, register durations, and the effect of measurement error.

probability density functions of the error term as the difference between the registered and reported durations. First, we can see how a substantial proportion of cases are relatively unaffected by ME. This is reflected by the points lying around the diagonal line of the scatterplot, where survey and register durations are equal. In particular, 162 subjects, 42% of the sample, reported durations within ± 15 days of what was captured in the register. This pattern is also manifested by the dotted line depicting differences between registered and reported durations in the density plot, which shows a majority of cases for which the observed ME appears to be well approximated by a Normal distribution centred around zero, as could be expected from classical additive ME. However, the same density function also shows that for a considerable proportion of the sample, durations have been markedly shortened, together with a few other spells reported to be artificially long. For these cases, there does not seem to be a particular relationship between Y^* and Y since – except for the cases subject to classical additive ME – they are roughly uniformly distributed across the window of observation (as depicted by the dotted line).

The complexity of the ME seen here can be regarded as the outcome of the different ME mechanisms affecting the retrospective report of work histories (presented in Figure 1). In the light of the sample restrictions — namely the selection of spells for which the respondents reported correctly to be unemployed — we can differentiate two ME mechanisms: 1) spells where the difference between the registered and reported durations are distributed around zero, which could be considered to be due to problems of misdating; and 2) spells that are spread across the window of observation, which could be due to the problems of omission or false positives (case 2 in Figure 1), or to problems of overreported spells (case 3 in Figure 1).

4.2. *Impact of Measurement Error in an Exponential Model*

To assess the consequences of using durations of unemployment prone to this type of ME, we now compare the results obtained for two accelerated-life exponential models, one using durations from the register as the response variable (the true model) and the other relying on the reported durations (the naïve model) – see Pina-Sánchez et al. (2013) for a detailed review of the impact of this type of ME in different types of event history analysis models. An exponential model is used instead of other commonly used Weibull or Cox specifications (Kettunen 1997; Lancaster 1979; Pyy-Martikainen and Rendtel 2009) for reasons of parsimony. Given the monotonic increase of the cumulative hazard functions for the reported and registered durations (Figure 3) and the complexity of the adjustments carried out in the following sections, it was deemed preferable to use the simplest plausible model specification. We include the same set of explanatory variables, *age*, work *experience*, and their interaction term in these two models. These variables could be considered nondifferential (see Table 1) with respect to the ME observed here, since the Pearson correlation coefficients between the ME (defined as $V = Y^* - Y$) with *age* and *experience* were 0.07 and -0.01.

To facilitate comparisons with the adjustments presented in the following section, the true and naïve models are estimated using Bayesian methods. We specify a model with a hierarchical dependence structure that lends itself to straightforward estimation in software such as WinBUGS (Lunn et al. 2000), JAGS (Plummer et al. 2006), or Stan

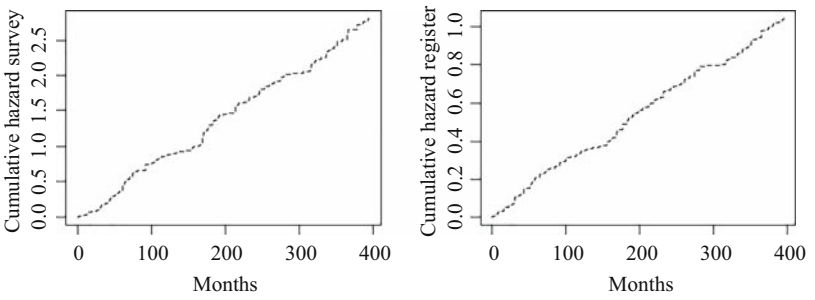

Fig. 3. Cumulative hazard functions for the reported and registered spells.

(Carpenter et al. 2017). All of these Bayesian packages are based on the Markov chain Monte Carlo (MCMC) approach (Geman and Geman 1984) that may be used when the full conditional posterior distributions of all unobservables are known distributions that are simple to draw from "directly".

The joint distribution for the true model can be formally expressed as

$$p(Y, \beta|X) = p(\beta)\prod_i p(Y_i|\mu_i), \tag{1}$$

where X represents the three explanatory variables, *age*, *exp*, and their interaction term *ae*, and μ_i is the expected value of Y_i. We assume that Y_i is exponentially distributed where we model μ_i using the link function $X_i\beta = \log(\mu_i)$. To complete the specification of the model we give diffuse priors to the regression estimates included in μ. For our 381 observations, we expect the likelihood to overwhelm the prior and we assume, a priori, that $\beta \sim N_4(0, 100^2 I_4)$. To estimate the naïve model, we only have to substitute Y for Y^* in Equation 1. Results from both the true and naïve models are shown in Table 3.

Although all the estimated coefficients have the same sign in both the naïve and true models, the effect of ME is clearly reflected by substantial attenuation in the estimates for the model using survey data. In particular, the effects of *age*, *experience* and their interaction on the durations of spells of unemployment are at least 90% smaller in the naïve model than in the true model. This has been measured using the relative bias (R.BIAS) for coefficient $\hat{\beta}_k^{(\text{survey})}$ relative to the true regression coefficient, $\hat{\beta}_k^{(\text{register})}$, $R.BIAS = 100\left|\hat{\beta}_k^{(\text{survey})} - \hat{\beta}_k^{(\text{register})}\right| / \left|\hat{\beta}_r\right|$, for $k = 1, 2, 3, 4$.

Table 3. Results of the exponential true and naive models and the impact of measurement error in terms of R.BIAS and R.RMSE.*

	Register	Survey	R.BIAS	R.RMSE
constant	9.10 (1.36)	5.12 (1.15)	43.7%	204.6%
age	−0.088 (0.038)	−0.001 (0.032)	89.9%	143.9%
experience	−1.39 (0.50)	−0.13 (0.42)	90.6%	171.1%
age × exp	0.038 (0.014)	0.002 (0.012)	94.7%	163.7%

*Results for all the models presented in this article are calculated from the posterior distributions formed from two chains of 400,000 iterations after the first 10,000 from each chain are burnt-in.
**Posterior standard deviations are shown within brackets.

The R.BIAS is useful for making comparisons between explanatory variables using different scales. In addition, in order to take into account the impact on the precision of the regression coefficients, we use the relative root mean squared error (R.RMSE), which is the root mean squared error of a regression coefficient obtained in the naïve model, $RMSE\left(\hat{\beta}_k^{(\text{survey})}\right) = \sqrt{SD\left(\hat{\beta}_k^{(\text{survey})}\right)^2 + (BIAS)^2}$, over that of the same coefficient in the true model, $R.RMSE = 100\frac{\left|RMSE\left(\hat{\beta}_k^{(\text{survey})}\right) - SD\left(\hat{\beta}_k^{(\text{register})}\right)\right|}{SD\left(\hat{\beta}_k^{(\text{register})}\right)}$.

In our study, the naïve model underestimates the standard deviations of all the regression coefficients. However, due to the attenuation of those estimates, the posteriors for the naïve model cover zero except for the constant term. The combination of bias and imprecision in the naïve model makes the impact of ME in terms of RMSE range from being doubled for the case of the constant term, to an increase of 43.9% (for *age*).

5. Adjustment

The great impact of the ME found in spells of unemployment derived from retrospectively reported work histories just shown is even more worrying if we take into account the difficulty of adjusting for such a complex process. Most of the standard methods designed for the adjustment of ME rely on rather simple assumptions regarding the behaviour of the ME. For example, methods such as SIMEX (Cook and Stefanski 1994), or the methods of moments (Fuller 1987) tend to assume that the error is classical (see Table 1), whereas other methods that can easily account for systematic ME, such as multiple imputation (Brownstone and Valletta 1996; Cole et al. 2006; Freedman et al. 2008; Messer and Natarajan 2008; Peytchev 2012; Rubin 1987) or regression calibration (Carroll and Stefanski 1990; Freedman et al. 2008; Glesjer 1990; Messer and Natarajan 2008; Veronesi et al. 2011; Wang et al. 1997) assume that the ME process is homoscedastic.

The problem of ME that changes in size according to the true value is typical in the retrospective report of start or end of spells. Several studies in the literature have investigated the adjustment of such types of errors using multiplicative models (Augustin 1999; Biewen et al. 2008; Dumangane 2007; Glewwe 2007; Holt et al. 2011; Pickles et al. 1996; Pickles et al. 1998; Pina-Sánchez 2016; Skinner and Humphreys 1999). However, as seen in the previous section, the type of ME observed in spells of unemployment derived from retrospectively reported work histories cannot be approximated using a multiplicative model. In particular, we have seen how, in addition to typical problems of spells being misdated, we should expect more serious problems of omission/overreport and misclassification of spells, which can generate cases where the observed and true durations are unrelated.

To adjust for such complex types of ME, we rely on the flexibility of the Bayesian approach. The possibility of specifying the ME freely to map the error-generating mechanism adequately is the key element that gives Bayesian methods an advantage over the previously discussed methods in terms of the flexibility and applicability in treating missing data (Rubin 1996). See for example, Clayton (1992) for a general framework; Richardson and Gilks (1993) demonstrated how the missing data allows for treatment of a wide variety of ME; Dellaportas and Stephens (1995) deal with both Berkson and classical

errors-in-variables ME for regressors; Butts (2003) treat ME in perceptions of social interaction; Ghilagaber and Koskinen (2009) correct for ME stemming from retrospectively defined covariates.

In what follows, we explore the possibility of extending the naïve model presented using a mixture measurement model with the aim of differentiating between two ME processes. The first process will reflect cases being either correctly reported or misdated. The second will deal with spells that result from misclassification, omission or overreporting of events, and for which the observed duration cannot offer much, if any, meaningful information about the true duration. In order to inform the mixture model about the allocation of cases to each of these processes, we explore two approaches: one relies on having access to a validation subsample, while the other relies on specific prior information.

5.1. Adjustments Relying on Validation Subsamples

For the first of these adjustments, we use random validation subsamples of 20% and 40% of the true durations captured in the original sample of 381 respondents. Compared to other studies seeking to adjust for ME, 20% is a relatively small subsample – for example Cole et al. (2006) used validation subsamples of 150 cases, accounting for 25% of the original sample – but large enough to make the model identifiable regardless of the configuration of the validation subsample. This was not always the case for smaller validation subsamples of 5% and 10%. A 20% validation subsample will, on average, only have 44 noncensored and 32 censored cases.

We present a general hierarchical mixture measurement model and then proceed to define the parameter structure and prior distributions. We let $\Lambda = \left(\beta, \theta, \sigma_1^2, \sigma_2^2, \pi\right)$ denote the collection of parameters of interest and introduce a latent variable $T = (T_i: i = 1, \ldots, 381)$. With prior distribution $p(\Lambda)$, the joint distribution of data and the unobservables is

$$p(Y^*, Y, \beta, \Lambda, T|X) = \prod_i p\left(Y_i^*|Y_i, \Lambda, T_i\right) p(T_i|X_i, Y_i, \Lambda) \tag{2}$$

$$\times \prod_i p(Y_i|X_i, \Lambda) \tag{3}$$

$$\times\ p(\Lambda), \tag{4}$$

where the model (3) for the true response is defined as an exponential distribution $p(Y_i|X_i, \Lambda) = p(Y_i|X_i, \beta)$ as before, but conditionally on a response Y_i, we define the ME distribution (2) for Y_i^* as a mixture of different ME models.

5.1.1. Mixture Measurement Model

The mixture model permits us to account for different types of ME simultaneously. We could, for example allow some cases to follow a multiplicative ME, while others follow an additive ME model. Here, we limit the case to two types of ME. For each respondent $i = 1, \ldots, 381$, we assume that they belong to one of two unobserved categories indicated by the latent variable $T_i \in \{0, 1\}$. If $T_i = 1$, we assume a standard additive ME

model $Y_i^* \big| \left[T_i = 1, Y_i, \sigma_1^2\right] \sim N\left(Y_i, \sigma_1^2\right)$. Conditional on $T_i = 0$, we assume that Y_i^* is completely unrelated to the true duration Y_i. Denoting this distribution $f\left(Y_i^*\right)$ and writing $\pi_i = \Pr(T_i = 1 | \pi, X_i, Y_i)$, the two processes correspond to a mixture

$$\pi_i \phi\left(Y_i^*; Y_i, \sigma_1^2\right) + (1 - \pi_i) f\left(Y_i^*\right),$$

where $\phi(\cdot;\ a,\ b)$ represents the pdf of a normal distribution with mean a and variance b.

The intuition is that additive ME might work for a subset of data but not for another subset. For a purely additive model, this heterogeneity would have to be accounted for entirely by the variance σ_1^2. A number of distributions for $f(\cdot)$ are conceivable and the shape of the distribution may contribute information and help discriminate between additive and random ME. To allow for $f(\cdot)$ to depend on for example X, would not alter the general structure of the model as long as the parameters of $f(\cdot | X)$ were distinct from β. Here we assume a normal distribution $f\left(Y_i^*\right) = \phi\left(Y_i^* \big| \theta, \sigma_2^*\right)$ which is convenient and can be interpreted as a variance decomposition of the ME. The conditional predictive distribution for Y_i^* with unknown T_i simplifies to

$$Y_i^* \big| \left[\pi_i, Y_i, \theta, \sigma_1^2, \sigma_2^2\right] \sim N\left(\pi_i Y_i + (1 - \pi_i)\theta, \pi_i^2 \sigma_1^2 + (1 - \pi_i)^2 \sigma_2^2\right).$$

Since respondents cannot report negative durations, Y_i^* can be truncated to the left in 0. Here we chose to relax this constraint and permit Y_i^* to have support $\mathbb{R}$. The reason for this is partly because the MCMC becomes less efficient with the truncation. Another consideration is that the truncation confounds the variance partition and makes interpretation of ME in terms of classical ME difficult. The distribution of the true value Y_i given Y_i^* and Λ, will be nonnegative even if Y_i^* is not truncated (see Section 7 Appendix for details).

5.1.2. Mixture Proportions

The latent variable, T, denoting the part of the mixture model to which cases are allocated, is set to follow a Bernoulli distribution $\Pr(T_i = 1 | Y,\ X,\ \Lambda) = \pi$, independently for all respondents conditional on the mixture proportion π. A priori π determines what proportion of cases follow the classical ME and what cases are unrelated. A posteriori, the predictive distribution for individual memberships also incorporates the information in the other variables. Like other latent variable models, it is possible to model π_i but we chose the more straightforward case.

5.1.3. Prior Distributions

To balance the evidence in data for the different ME processes and to adjust proportions, we need to pay careful attention to prior distributions. To reflect the structure of the model, we set prior distributions

$$p(\Lambda) = p(\pi) p(\beta)\, p\left(\theta, \sigma_1^2, \sigma_2^2\right),$$

assuming a priori independence between different blocks of parameters. As $\pi \in (0, 1)$, a convenient and common choice is $Beta(\zeta_1, \zeta_2)$, with standard reference priors being $\zeta_1 = \zeta_2 = 1/2$ (Jeffreys, 1946) or a uniform distribution $\zeta_1 = \zeta_2 = 1$. Here we chose the latter.

The analysis of the full register data was not very sensitive to the prior for β. With 20% or 40% validation samples, the scant information in data will, however, give more weight to the prior distribution. For regression parameters β, it is common in generalised linear models to use the prior distribution $\beta \sim N_p(0, \text{diag}(\lambda))$, for $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$. This may be interpreted as a prior that shrinks the coefficients towards the origin. We set a relatively conservative $\lambda = (100^2, 10^2, 10^2, 10^2)^{\mathrm{T}}$, where the variance λ_1 for β_{cons} reflects the greater variation in this parameter and prevents too great attenuation of the intercept. Other alternatives include the class of conjugate families (Chen and Ibrahim 2003), or other forms of reference priors – for example, Ibrahim and Laud (1991) – that can account for the difference in scale and correlation between covariates.

As the ME model attempts to parse out the relative contribution of variance in $Y* - Y$, from the two kinds of ME, the model is potentially sensitive to the prior specifications for $p\left(\theta, \sigma_1^2, \sigma_2^2\right)$. Since we do not have any theory to guide us and given the structure of the mixture model, we assume that $\left(\theta, \sigma_2^2\right)$ are independent of σ_1^2 a priori. Here, we consider two different kinds of prior distributions.

The first type is a standard conjugate prior

$$\frac{1}{\sigma_1^2} \sim Ga(\alpha_1, \gamma_1),$$

and for the parameters of f we chose a standard normal-inverse-gamma prior

$$\theta \big| \sigma_2^2 \sim N\left(\theta_0, \sigma_2^2/n_0\right) \quad \frac{1}{\sigma_2^2} \sim Ga(\alpha_2, \gamma_2),$$

for hyper parameters $\alpha_1,\ \alpha_2 > 2,\ \gamma_1,\ \gamma_2 > 0$, and $n_0 > 0$. To make sure that both the variances and the precisions have proper distributions with the first two moments finite, we set $\alpha_1 = \alpha_2 = 3$. This is to ensure that the conditional posteriors are well defined for the case when $\sum_i T_i = 0$ or $\sum_i T_i = 381$, that is, when the model allocates all observations to either one of the two latent classes. Throughout, we will set $n_0 = 1$ and $\theta_0 = 187$, which is exactly the centre of the range of observable values on Y_i^*. Because of the (conditional) conjugacy, updates of α_1,α_2, γ_1,γ_2, and θ are efficient. Similarly, updates of unobserved Y_i are straightforward given the result in the Section 7 Appendix.

To relax the dependence between θ and σ_2^2 somewhat, and make $f()$ slightly more robust to violations of normality, we also consider a prior where θ and σ_2^2 are independent a priori. More specifically, we set the prior for $\theta \sim N(\theta_0, \tau^2)$, $\theta_0 = 187$, and independently thereof $\sigma_s^2 \sim U(1, a_s)$, for some choice of upper bound a_s, $s = 1, 2$ (Gelman et al. 2006, 520).

5.2. *Performance of the Adjustment*

To investigate the performance of the adjustment, we assess the sampling error associated with the choice of validation subsample, as well as the sensitivity to prior specifications. We estimate the model from 50 samples of 20% and 40% of the register data for prior specifications, as in Table 4.

5.2.1. 20% Validation Sample

A comparison of the posteriors for β with the naïve analysis and the gold standard is provided in Figure 4 that provides the 95% credibility intervals (CI) for the adjustments

Table 4. Specifications for hyper parameters for σ_1^2 and σ_2^2.

	γ_1	γ_2	$E(\sigma_1^2)$	$E(\sigma_2^2)$	$V(\sigma_1^2)$	$V(\sigma_2^2)$
Prior I	5	10	2.5	5	6.25	25
Prior II	10	100	5	50	25	2500
Prior III	100	400	50	200	2500	40000

under priors I, II, and II, as well as the CIs and means for the validation sample only and register analyses. While the posteriors for π, σ_1^2, σ_2^2, and θ clearly are sensitive to the prior specifications, the posteriors for the regression parameters β appear robust to the priors for the ME. With as few as 76 validation cases, it appears that we get valid inference for both β and T and there seems too be a very weak dependence on the actual sample judging by the small variation across samples. The posterior means for the adjustments seem to be close to those obtained using the validation sample only, but the CIs for the latter are much wider than for the adjustments. Using the survey only yields less uncertainty and narrower CIs (as shown in Table 3) than the adjustment, but the bias of the survey results means that the CIs do not cover the true posterior mean for any parameter.

All regression coefficients for the adjustment are attenuated and the CIs are somewhat wider. For β the adjustment CIs are close to identical for Priors I, II, and III. There is a

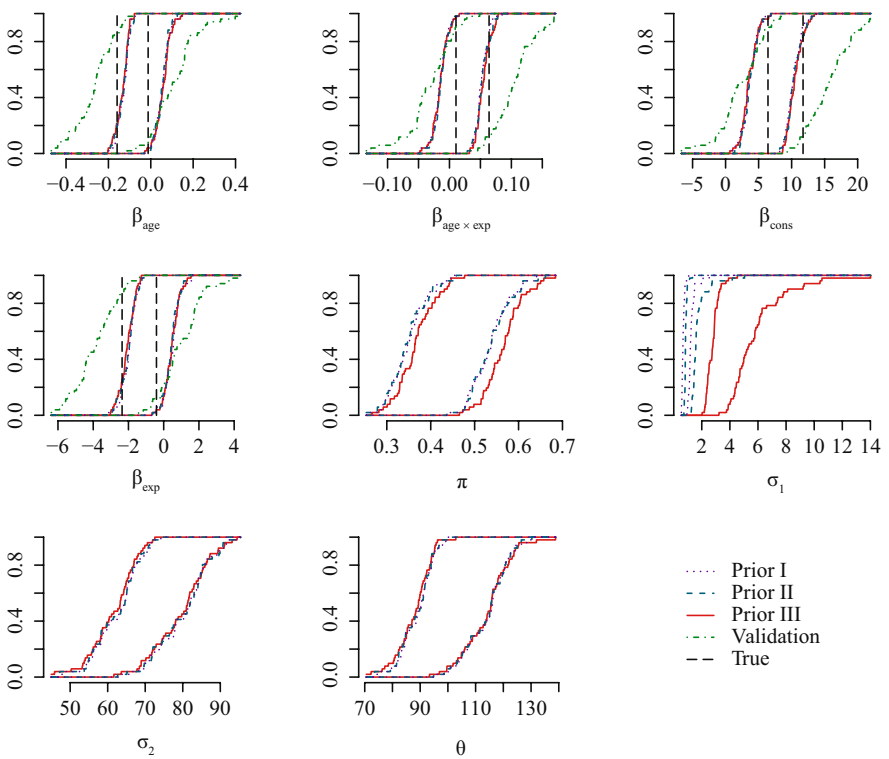

Fig. 4. 20% validation samples: CDF of Credibility intervals under prior specifications I, II, III, for 50 replicates, compared to the naïve estimates and the validation sample only.

healthy overlap between the adjustment CIs and the true CIs, but the interval lengths for the latter are shorter. The relative bias for *age* ranges between 5 and 129, with a mean of 58; for the constant, the relative bias ranges between 1 and 46 with a mean of 23.

We may note that on balance, the first part of the mixture model $T_i = 1$ makes a relatively small contribution in the adjustment. The fact that the model can differentiate between these cases and those much more seriously affected by ME is critical for the success of the adjustment. Using the 20% validation subsample, the mixture model estimates that the proportion of cases set to $T_I = 1$ (π) is generally in the range 30% to 60% which covers 46%, the proportion of cases in the validation subsample where reported durations lay within ± 15 days of what was captured in the register.

While the inference for β seems robust to the prior specification using the tight coupling of θ and σ_2^2 of the conjugate prior, we also fitted the adjustment model using the second kind of prior with $a_1 = 50$ and $a_2 = 500$ (Prior IV). Again, the posterior means and standard deviations are not affected by the prior specification for the ME. Prior IV partitions the variance of the ME more clearly but is very similar to Priors I and II (Figure 5).

5.2.2. 40% Validation Sample

Increasing the size of the validation subsample to 40% yields more information both for the regression parameters β and the ME process. We re-estimate the model under Prior II

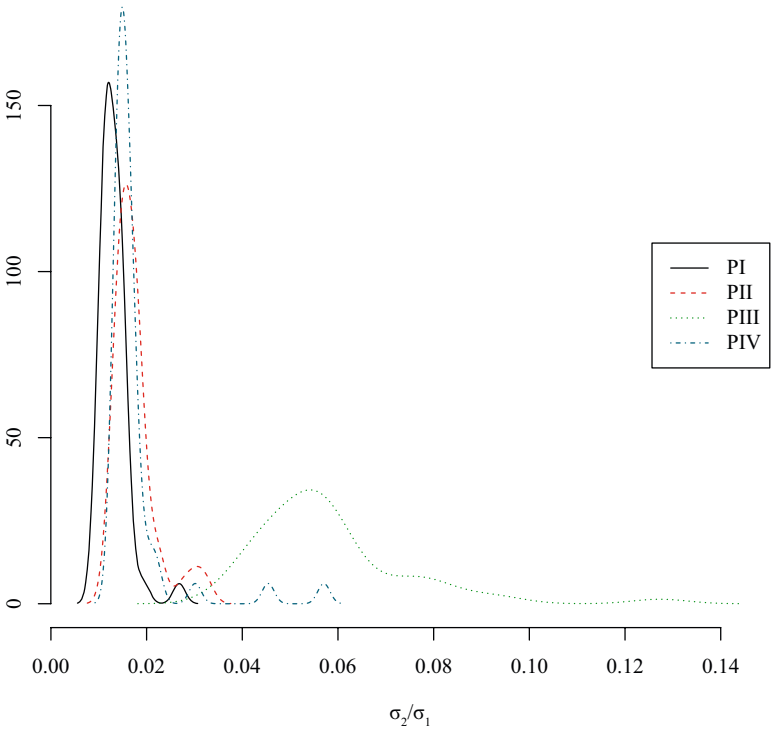

Fig. 5. Comparison of posterior ratio between variance contributions of ME different types of priors across 20% validation samples.

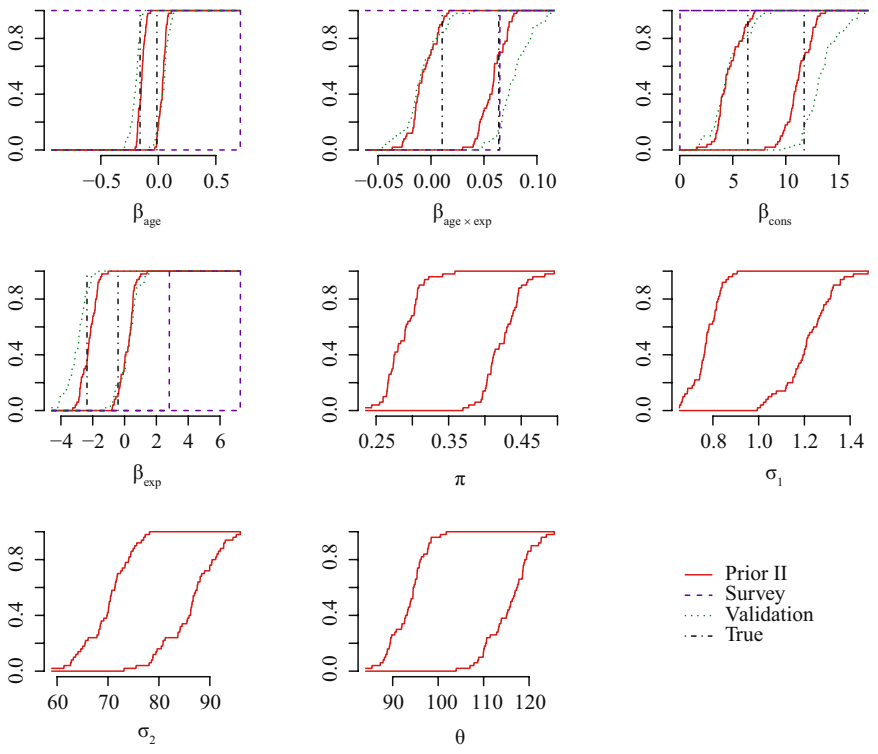

Fig. 6. 40% validation samples: CDF of Credibility intervals under prior specification II, for 50 replicates, compared to the naive estimates and the validation sample only.

(given that the inference for 20% was robust to choice of prior). The posterior means for β are still somewhat attenuated, but the CIs (Figure 6) are now very close to the intervals for the register data. This improvement is not matched by the validation-only analyses. For example, the average interval length for *age* for the adjustment is 0.18 and for validation only it is 0.24.

5.3. Adjustments Without Validation Sample

Without a validation subsample, the previous model would have been unidentifiable since it would be impossible, using the survey data alone, to determine the probability of cases belonging to each part of the mixture model. However, the requirement of having access to validation data can be relaxed using more informative priors. In particular, we obtain results by assuming that the probability of cases falling within each of the two parts of the mixture model is fixed. Prior distributions become a very useful tool in the presence of models that cannot be identified – as is often the case when dealing with ME problems. *"One intuitive way of thinking about Bayesian inference in the absence of parameter identifiability is that the prior distribution plays more of a role than usual in determining the posterior belief about the parameters having seen the data"* (Gustafson 2003, 64). However, this greater reliance on the priors implies a reduced role for the data and, in consequence, an increased possibility of incurring in model misspecification.

The more informed the researcher is about the value of the model parameters, the lower the probability of misspecification will be. Hence, studies designed to assess the presence of ME are essential to carrying out adequate adjustments. A number of studies have explored the problems of ME affecting retrospectively reported work histories (Biemer 2011; Kreuter et al. 2010; Manzoni et al. 2011 and 2010; Poterba and Summers 1984 and 1995; Pyy-Martikainen and Rendtel 2009). However, we are only aware of one study (Pyy-Martikainen and Rendtel 2009) using a register as validation data to assess the different types of ME found in these types of questions, and even here we should note some important differences from our study. Pyy-Martikainen and Rendtel (2009) looked at a representative sample of the Finnish population, studied spells reported over a period of five years, and the question used only required respondents to report the work status experienced in each month of the year. We, on the other hand, study a sample composed of Swedish jobseekers, observed for a period of one year, and responding to a question where every spell needed to be identified in chronological ordered and dated.

These differences in terms of the sample composition, window of observation, and question format, are so important that it would probably be unwise to borrow specific point estimates from Pyy-Martikainen and Rendtel (2009) for our adjustment. However, we can take the overall lower prevalence of ME found in Pyy-Martikainen and Rendtel (2009) into consideration as one of the scenarios that we explore. In particular, we carry out adjustments ranging from $\pi = 0.5$ – a rounded estimate of the proportion of reported durations that lay within ± 15 days of the true ones – to $\pi = 0.7$ – an estimate that assumes a lower prevalence of cases being misclassified, omitted or overrepresented. We assume Prior IV for the ME process.

Judging from Figure 7, inference for β is little affected by our choice of π but the added uncertainty of not having a validation sample is reflected in spread in the posteriors. The bias (relative to the register means) increases for all parameters as π increases. This is more clear in Table 5, and the bias for the two extremes are compared in Figure 8.

We see that the models explored succeeded in reducing the bias found in the naïve analysis (Figure 8), while some of the added uncertainty is reflected in inflated standard deviations. The model assuming $\pi = 0.5$ performed better at reducing the R.BIAS and

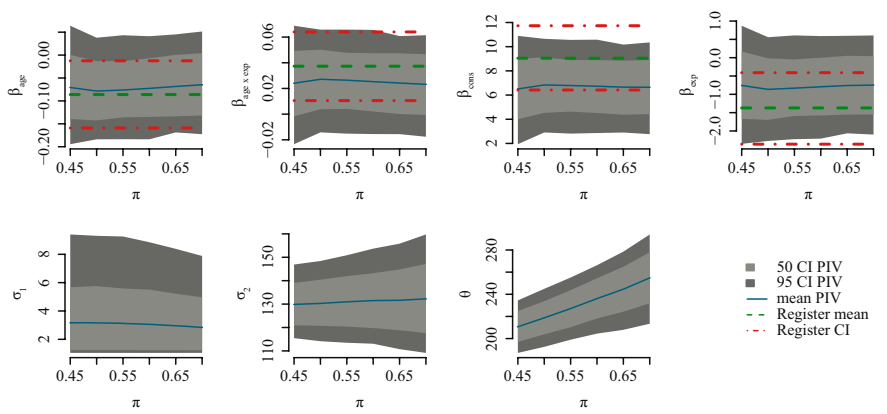

Fig. 7. No validation samples: Credibility intervals (50% and 95%) under prior specification IV against proportion π.

Table 5. Adjustment using mixture models with fixed proportions.*

	Register	Survey	$\pi = .5$	$\pi = .7$
constant	9.10 (1.36)	5.12 (1.15)	6.85 (1.98)	6.64 (1.94)
age	−0.088 (0.038)	−0.001 (0.032)	−0.079 (0.056)	−0.065 (0.058)
experience	−1.39 (0.50)	−0.13 (0.42)	−0.87 (0.72)	−0.75 (0.70)
age × exp	0.038 (0.014)	0.002 (0.012)	0.027 (0.020)	0.023 (0.020)
σ_1			3.13 (2.22)	2.80 (1.83)
σ_2			130.4 (8.7)	132.2 (12.9)
θ			218.9 (13.0)	254.5 (19.8)

*Posterior standard deviations are shown in brackets.

R.RMSE. The R.RMSE for the constant is reduced from 204% to 120%, and from 144% to 49% for *age*.

It seems that a value of π closer to what can be observed in the sample helps in the reduction of the R.BIAS observed in the naïve model. However, the lower the value of π the higher the proportion of cases being treated by the second part of the mixture model. Finally, if we compare the adjustments based on 20% and 40% validation subsamples, they generally have slightly higher R.BIAS and R.RMSE on average than models relying on a fixed $\pi \in (0.5, 0.7)$, but with a range that covers those for the latter. For example, R.BIAS for *age* for the 20% adjustment is between 4.7 and 129 with an average of 58; and the 40% adjustment is between 0.3 and 124 with an average of 40. R.RMSES for *age* for the 20% adjustment is between 14.9 and 215 with an average of 85; and the 40% adjustment is between 11 and 205 with an average of 58.

6. Discussion

As has been noted by different authors (Augustin 1999; Jäckle 2008; Pyy-Martikainen and Rendtel 2009), the understanding of and adjustment for ME in longitudinal data remains an understudied area. *"Despite the recognition of the existence of measurement errors in*

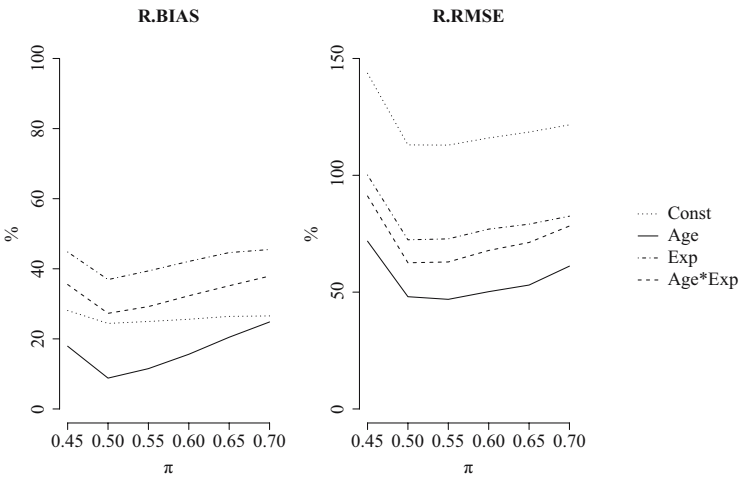

Fig. 8. Effectiveness of the adjustment using a mixture model with fixed π.

survey-based data on event histories, little is known about their effects on an event history analysis." (Pyy-Martikainen and Rendtel 2009, 140). Our intention here has been to contribute to this topic by studying the prevalence, impact, and adjustments of the type of ME observed in spells of unemployment derived from retrospectively collected work histories using register data. From this analysis, we would like to underline three main points, each being a consequence of the previous one:

1) Unlike the typical problems of time displacement found in reports of the starts or ends of specific spells, here we have noted that reports of work histories can also be prone to ME in the form of misclassification omission and overreporting of spells. Furthermore, these different types of ME can occur simultaneously, generating complex ME patterns.
2) Misclassification, omission and overreporting of spells are more problematic forms of ME than misdated spells because they tend to result in observed durations that are unrelated to the true values, hence endangering both the validity of these measures and the analyses based on them. In particular, we have shown how, when spells of unemployment derived from retrospectively reported work histories are used as the response variable of an event history model, its regression coefficients are severely attenuated by up to 90% of their true value.
3) Because of the complexity of the ME processes studied, standard adjustment methods are inadequate. In particular, it is important to note that the vast majority of studies aiming to adjust for the ME found in retrospective reports of dates have been based on elaborations of multiplicative ME, which are inadequate in the presence of spells being omitted, overreported, or misclassified.

Here, we have proposed to adjust for different ME mechanisms simultaneously through the use of a Bayesian mixture measurement model. Specifically, the mixture model allows us to differentiate between spells that are correctly reported or affected by minor problems of misdating, and those for which the observed durations are not related to their true value. To inform the mixture model about the proportion of cases that needed to be predicted according to each of the mechanisms, we explored two approaches.

First, we assumed that the researcher could have access to a validation subsample. The method performs well for 40% validation sample, but also works for 20% validation sample. Here, 20% means only 76 cases from the register (32 on average of which are censored), but even such a small figure is often inaccessible to most researchers due to reasons of confidentiality. As an alternative to using validation data, we demonstrate the use of fixing, a priori, the proportion of subjects in one of the two ME processes when no validation subsample is available. These adjustments were even more effective. For a model where the probability of being in the first part of the mixture model is fixed at 0.5 – reflecting the percentage of spells observed to be correctly reported or simply mildly misdated – we obtained average reductions of the R.BIAS of 65%, whereas fixing that probability at 0.7 showed average reductions of 54%. In addition, these models fixing the probabilities of the mixture model at 0.5 and 0.7 reduced the R.RMSE by 41% and 47%, respectively.

The adjustment methods presented here are potentially useful strategies to reduce the impact of the types of ME that can be expected in the report of life-course histories. The

adjustments have been kept relatively simple and could be implemented by other researchers using data that are similarly prone to combinations of different types of ME. It is, for example, straightforward to model the mixture proportions as a function of observables and to relax assumptions for the form of unrelated ME. The question remains as to how to choose the necessary priors to make the model identifiable when no validation data is available. Findings from studies analysing the prevalence of ME in similar survey questions could help to inform those decisions, but even in the context of a well-analysed question, a sensitivity analysis, or the comparison of results obtained from the use of different priors, is the most prudent solution. For example, future studies specifying durations of unemployment stemming from similar retrospective questions to the one analysed here could use our approach to explore the robustness of their findings when the percentage of spells taken to be misclassified, omitted, or overreported grows from 0% to 20% and 40%. Future work is needed to explore to what extent replicate data from the posterior predictive distributions can be used to assess what ME has best fit to observed data (Gilks et al. 1996).

The quality of the adjustment could also be improved using the, nowadays, more frequently coded paradata. For example, the probability of cases being considered by each part of the mixture model could be made conditional on certain factors known to be associated with the misclassification, omission or overreporting of spells of unemployment, such as whether the interview was taken on the phone rather than face-to-face (Mathiowetz and Duncan 1988). Such adjustments could be further improved if the researcher has access to key socio-demographic characteristics of the respondent. For example, it has been consistently detected that individuals less engaged in the labour market are more prone to omit spells of unemployment (Bound et al. 2001; Jürges 2007; Levine 1993; Morgenstern and Barrett 1974; Paull 2002).

We would also like to add a note of comfort regarding the high levels of both the prevalence and the impact of the ME analysed here, since there are reasons to believe that they might be unusually high. First, the sample under study here is not representative of the Swedish population, as it is composed of strictly unemployed subjects, who, for reasons of social desirability might have a higher tendency to omit their spells of unemployment. Second, the register of unemployment is not infallible. We detected coding errors in the form of nonsensical dates, for example spells that started before the previous spell had ended, or others dated the 32nd day of a month. These cases were dropped from the sample, but other coding errors in the register might have gone undetected, which should make us aware that a proportion of the differences between the register and the survey were actually reflecting ME in the former and not in the latter. Finally, the format of the question used is cognitively quite demanding. Nowadays, longitudinal surveys like the European Union Statistics on Income and Living Conditions (EU-SILC), or the Swedish Level of Living Survey (LNU), use questions where respondents are only asked to report their work status in each of the months of the previous year. These questions will fail to detect transitions shorter than a month, but since they are easier to answer we could also expect lower levels of ME than in questions where all work-related spells are asked to be reported and dated with day-level detail.

The different format of these questions affects the level of measurement of the variables to be retrieved from them, and with that, the modelling strategies to be used. Rather than

obtaining duration data amenable to parametric event history model specifications – like the exponential model presented here – person-period categorical data is obtained, which is more suitably specified using nonparametric models (Box-Steffensmeier and Jones 2004). The details of the adjustment presented here would not be directly applicable in those instances. Instead, we would direct the interested reader to relatively recent studies from Manzoni et al. (2010) and Biemer (2011) implementing adjustments based on latent Markov models and repeated measures.

7. Appendix – Conditional Predictive Distribution for True Values

For the un-truncated normal distribution, the location and scales are independent and normal conjugacy applies. This makes it straightforward to model θ, σ_2^2, and σ_1^2 such that data are able to determine class memberships (T_i: $i = 1, \ldots, 381$). The conditional posterior predictive distribution of the true value Y_i given Y_i^* and Λ, is a normal distribution truncated to the left in 0. Note that this only holds when Y_i^* itself is not truncated.

Let $\eta = \pi Y + (1 - \pi)\theta$, $\tau = \frac{1}{2(\pi_i^2\sigma_1^2 + (1-\pi_i)^2\sigma_2^2)}$, and set $z = Y^*$, then $y|[z, \Lambda]$ has pdf

$$p(y|z, \Lambda) = \frac{p(z|y, \Lambda)p(y|\Lambda)}{\int p(z|y, \Lambda)p(y|\Lambda)\mathrm{d}z} = \frac{\exp\{-\tau(z - \eta)^2 - y/\mu\}1(y > 0)}{\int \exp\{-\tau(z - \eta)^2 - y/\mu\}1(y > 0)\mathrm{d}y}.$$

The integrand in the denominator can be written

$$\exp\{-\tau(z - \eta)^2 - y/\mu\}\mathbf{1}(y > 0) = h(z, \Lambda)g(y; z, \Lambda)$$

where $h(z, \Lambda)$ is a only a function of z and Λ, and

$$\begin{aligned} g(y; z, \Lambda) &= \exp\left\{-\tau\pi^2\left(y^2 - y\lambda + \frac{ts}{\pi^2}\right)\right\}\mathbf{1}(y > 0) \\ &= \exp\left\{-\tau\pi^2(y - \lambda)^2 - \tau\pi^2\left(\lambda^2 - \frac{ts}{\pi^2}\right)\right\}\mathbf{1}(y > 0) \\ &= (\pi\tau)^{-1/2}\phi(y; \lambda, \upsilon^2)c(z, \theta, \mu)\mathbf{1}(y > 0) \end{aligned}$$

Where $s = z + (1 - \pi)\theta$,

$$\lambda = \frac{\pi t - \frac{1}{\mu\tau}}{\pi^2}, \quad \upsilon^2 = \frac{1}{2\pi^2\tau}, \quad t = z - (1 - \pi)\theta,$$

and

$$c(z, \theta, \mu) = e^{-\tau\pi^2(\lambda^2 - ts/\pi^2)}.$$

Since $c(z, \theta, \mu)$ is a function of z, θ, and μ, only, and $h(z, \Lambda)$ is a only a function of z and Λ, we have

$$p(y|z, \Lambda) = \frac{\phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)}{\int \phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)\mathrm{d}y} = \frac{\phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)}{\Phi(-\lambda/\upsilon)}.$$

We recognise this as a normal distribution $N(\lambda, \nu^2)$, truncated in the left at 0. Consequently, even when f () is inconsistent with data in the sense that Y_i^* may take negative values, the conditional predictive distribution for the true values Y_i is still consistent with data.

Consider now the case when the observations with error Y^* are themselves truncated to the left in 0. For the normal distribution $\phi(\cdot; a, b)/\Phi(-a/b)$ truncated to the left in 0, the variance is a function of a (for example, for fixed b, setting $a < 0$ increases the variance). For our data, the truncation results in θ becoming negative so that the variance in $Y_i^* \big| \left[\theta, \sigma_2^2, T_i = 0\right]$ is determined by the right tail of the truncated distribution rather than the variance of the distribution. Negative mean θ and the dependence of $V\left(Y_i^* \big| \theta, \sigma_2^2, T_i = 0\right)$ on both θ and σ_2^2 makes the interpretation in terms of classical ME difficult and sampling using MCMC inefficient.

8. References

Augustin, T. 1999. Correcting for Measurement Error in Parametric Duration Models By Quasi-likelihood. Technical Report, Max Plank Institute.

Berkson, J. 1950. "Are There Two Regressions?" *Journal of the American Statistical Association* 45(250): 164–180. Doi: https://doi.org/10.2307/2280676.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Error*. Wiley.

Biewen, E., S. Nolte, and M. Rosemann. 2008. "Perturbation by Multiplicative Noise and the Simulation Extrapolation Method." *Advances in Statistical Analysis* 92: 375–389. Doi: https://doi.org/10.1007/s10182-008-0089-7.

Black, D.A., M.C. Berger, and S.A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95(451): 739–748. Doi: https://doi.org/10.2307/2669454.

Bound, J., C. Brown, and N.A. Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, edited by J. Heckman and E. Leamer. Vol. 5: 3705–3843. New York: Elsevier.

Box-Steffensmeier, J.M. and B.S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge: Cambridge University Press.

Bradburn, N.M., J. Huttenlocher, and L. Hedges. 1994. "Telescoping and Temporal Memory." In *Autobiographical Memory and The Validity of Retrospective Reports*, edited by N. Schwarz and S. Seymour, 203–215. New York: Springer.

Brownstone, D. and R.G. Valletta. 1996. "Modelling Earnings Measurement Error: A Multiple Imputation Approach." *The Review of Economics and Statistics* 78(4): 705–717. Doi: https://doi.org/10.2307/2109957.

Butts, C.T. 2003. "Network Inference, Error, and Informant (in) Accuracy: A Bayesian Approach." *Social Networks* 25(2): 103–140. Doi: https://doi.org/10.1016/S0378-8733(02)00038-2.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M.A. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1): 1–32. Doi: https://doi.org/10.18637/jss.v076.i01.

Carroll, R.J. and L.A. Stefanski. 1990. "Approximate Quasilikelihood Estimation in Models with Surrogate Predictors." *Journal of the American Statistical Association* 91: 242–250. Doi: https://doi.org/10.2307/2290000.

Chen, M.H. and J.G. Ibrahim. 2003. "Conjugate Priors for Generalized Linear Models." *Statistica Sinica* 13: 461–476. Available at: http://www3.stat.sinica.edu.tw/statistica/oldpdf/a13n212.pdf (accessed February 2019).

Clayton, D.G. 1992. "Models for the Analysis of Cohort and Case-control Studies with Inaccurately Measured Exposures." *Statistical Models for Longitudinal Studies of Health*: 301–331.

Cole, S., H. Chu, and S. Greenland. 2006. "Multiple-imputation for Measurement-error Correction." *International Journal of Epidemiology* 35: 1074–1081. Doi: https://doi.org/10.1093/ije/dyl097.

Cook, J. and L. Stefanski. 1994. "A Simulation Extrapolation Method for Parametric Measurement Error Models." *Journal of the American Statistical Association* 89: 1314–1328. Doi: https://doi.org/10.2307/2290994.

Crowder, R.G. 1976. "The Interference Theory of Forgetting in Long-term Memory." In *Principles of Learning and Memory*, edited by R.G. Crowder. Oxford: Lawrence Erlbaum.

Dellaportas, P. and D.A. Stephens. 1995. "Bayesian Analysis of Errors-in-variables Regression Models." *Biometrics* 51(3): 1085–1095. Doi: https://doi.org/10.2307/2533007.

Dumangane, M. 2007. Measurement error bias reduction in unemployment durations. Technical report, CEMMAP. Doi: https://doi.org/10.1920/wp.cem.2006.0306.

Freedman, L.S., D. Midthune, R.J. Carroll, and V. Kipnis. 2008. "A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error in Regression." *Statistics in Medicine* 27: 5195–5216. Doi: https://doi.org/10.1002/sim.3361.

Fuller, W. 1987. *Measurement Error Models*. New York: John Wiley and Sons.

Gelman, A. et al. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1(3): 515–534. Doi: https://doi.org/10.1214/06-BA117A.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions On Pattern Analysis and Machine Intelligence* 6: 721–741. Doi: https://doi.org/10.1109/TPAMI.1984.4767596.

Ghilagaber, G. and J. Koskinen. 2009. "Bayesian Adjustment of Anticipatory Covariates in the Analysis of Retrospective Data." *Mathematical Population Studies* 16(2): 105–130. Doi: https://doi.org/10.1080/08898480902790171.

Gilks, W., S. Richardson, and D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Glesjer, L. 1990. "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-variables Regression Models." In *Statistical Analysis of Error Measurement Models and Application*, edited by P. Brown and W. Fuller, 99–114. Providence: American Mathematics Society. Doi: https://doi.org/10.1090/conm/112.

Glewwe, P. 2007. "Measurement Error Bias in Estimates of Income and Income Growth Among the Poor: Analytical Results and a Correction Formula." *Economic Development and Cultural Change* 56: 163–189. Doi: https://doi.org/10.1086/520559.

Golub, A., B.D. Johnson, and E. Labouvie. 2000. "On Correcting Biases in Self-reports of Age at First Substance use with Repeated Cross-section Analysis." *Journal of Quantitative Criminology* 16: 45–68. Doi: https://doi.org/10.1023/A:1007573411129.

Gustafson, P. 2003. *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: Chapman and Hall.

Holt, D., J.W. McDonald, and C.J. Skinner. 2011. "The Effect of Measurement Error on Event History Analysis." In *Measurement Error in Surveys*, edited by P. Biemer, 665–685. New York: John Wiley.

Huttenlocher, J., L. Hedges, and V. Prohaska. 1988. "Hierarchical Organization in Ordered Domains: Estimating the Dates of Events." *Psychological Review* 95: 471–484.

Ibrahim, J.G. and P.W. Laud. 1991. "On Bayesian Analysis of Generalized Linear Models using Jeffreys's Prior." *Journal of the American Statistical Association* 86(416): 981–986. Doi: https://doi.org/10.1037/0033-295X.95.4.471.

Jäckle, A. 2008. Measurement error and data collection methods: Effects on estimates from event history data. Technical report, Institute for Social and Economic Research, ISER. Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2008-13.pdf (accessed February 2019).

Jeffreys, H. 1946. "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 24: 453–461. Doi: https://doi.org/10.1098/rspa.1946.0056.

Jenkins, S.P. and P. Lynn. 2005. *Improving Survey Measurement of Income and Employment, 2001–2003* (2nd ed.). UK Data Service. Doi: https://doi.org/10.5255/UKDA-SN-5157-1.

Johnson, E.O. and L. Schultz. 2005. "Forward Telescoping Bias in Reported Age of Onset: An Example From Cigarette Smoking." *International Journal of Methods in Psychiatric Research* 14: 119–129. Doi: https://doi.org/10.1002/mpr.2.

Jürges, H. 2007. "Unemployment, Life Satisfaction and Retrospective Error." *Journal of the Royal Statistical Society, Series A* 170(1): 43–61. Doi: https://doi.org/10.1111/j.1467-985X.2006.00441.x.

Kapteyn, A. and J.Y. Ypma. 2007. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data." *Journal of Labour Economics* 25(3): 513–551. Doi: https://doi.org/10.1086/513298.

Kettunen, J. 1997. "Education and Unemployment Duration." *Economics of Education Review* 16(2): 163–170. Doi: https://doi.org/10.1016/S0272-7757(96)00057-X.

Kreuter, F., G. Mller, and M. Trappman. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74(5): 880–906. Doi: https://doi.org/10.1093/poq/nfq060.

Lancaster, T. 1979. "Econometric Methods for the Duration of Unemployment." *Econometrica* 47(4): 939–956. Doi: https://doi.org/10.2307/1914140.

Levine, P. 1993. "CPS Contemporaneous and Retrospective Unemployment Compared." *Monthly Labor Review* 116: 33–39. Available at: https://heinonline.org/HOL/Page?handle=hein.journals/month116&div=89&g_sent=1&casa_token=pTR6IZj22XsAAAAA:xq9wIH0hhVt7hMJgw6ViXuW_gWKx8-EARBvTPW32LcaWEKxYad-v0O53OauyAW25tklO5TD6&collection=journals (accessed February 2019).

Lunn, D.J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. "Winbugs a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10: 325–337. Doi: https://doi.org/10.1023/A:1008929526011.

Manzoni, A., R. Luijkx, and R. Muffels. 2011. "Explaining Differences in Labour Market Transitions between Panel and Life-course Data in West-Germany." *Quality and Quantity* 45: 241–261. Doi: https://doi.org/10.1007/s11135-009-9292-1.

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

Mathiowetz, N. and G. Duncan. 1988. "Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment." *Journal of Business and Economic Statistics* 6(2): 221–229. Doi: https://doi.org/10.1080/07350015.1988.10509656.

Messer, K. and L. Natarajan. 2008. "Maximum Likelihood, Multiple Imputation and Regression Calibration for Measurement Error Adjustment." *Statistics in Medicine* 27(30): 6332–6350. Doi: https://doi.org/10.1002/sim.3458.

Morgenstern, R. and N. Barrett. 1974. "The Retrospective Bias in Unemployment Reporting By Sex, Race and Age." *Journal of the American Statistical Association* 69(346): 355–357. Doi: https://doi.org/10.2307/2285657.

Neter, J. and J. Waksberg. 1964. "A Study of Response Errors in Expenditures Data From Household Interviews." *Journal of the American Statistical Association* 59: 18–55. Doi: https://doi.org/10.1080/01621459.1964.10480699.

Neuhaus, J.M. 1999. "Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression." *Biometrika* 86(4): 843–855. Doi: https://doi.org/10.1093/biomet/86.4.843.

Novick, M.R. 1966. "The Axioms and Principal Results of Classical Test Theory." *Journal of Mathematical Psychology* 3: 1–18. Doi: https://doi.org/10.1016/0022-2496(66)90002-2.

Office for National Statistics. Social and Vital Statistics Division. 2006. General Household Survey, 2003–2004. [data collection]. 2nd Edition. UK Data Service. SN: 5150. Available at: http://doi.org/10.5255/UKDA-SN-5150-1.

Office for National Statistics. 2008. Social and Vital Statistics Division. ONS Omnibus Survey, April 2006. [data collection]. UK Data Service. SN: 5997. Available at: http://doi.org/10.5255/UKDA-SN-5997-1.

Paull, G. 2002. Biases in the reporting of labour market dynamics. Technical report, Institute for Fiscal Studies. Doi: https://doi.org/10.1920/wp.ifs.2002.0210.

Pavlopoulos, D. and J.K. Vermunt. 2015. Measuring temporary employment. do survey or register data tell the truth? Technical report, Vrije Universiteit Amsterdam. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14151-eng.htm (accessed February 2019).

Peytchev, A. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly* 76(2): 214–237. Doi: https://doi.org/10.1093/poq/nfr065.

Pickles, A., K. Pickering, E. Simonoff, J. Silberg, J. Meyer, and H. Maes. 1998. "Genetic Clocks and Soft Events: A Twin Model for Pubertal Development and Other Recalled Sequences of Developmental Milestones, Transitions, or Ages At Onset." *Behavior Genetics* 28: 243–253. Doi: https://doi.org/10.1023/A:102161522.

Pickles, A., K. Pickering, and C. Taylor. 1996. "Reconciling Recalled Dates of Developmental Milestones, Events and Transitions: A Mixed Generalized Linear Model with Random Mean and Variance Functions." *Journal of the Royal Statistical Society. Series* A1: 225–234. Doi: https://doi.org/10.2307/2983170.

Pina-Sánchez, J. 2016. "Adjustment of Recall Errors in Duration Data using Simex." *Advances in Methodology and Statistics* 12(1): 27–58. Available at: http://ibmi.mf.uni-lj.si/mz/2016/no-1/p3.pdf (accessed February 2019).

Pina-Sánchez, J., J. Koskinen, and I. Plewis. 2013. "Implications of Retrospective Measurement Error in Event History Analysis." *Metodología de Encuestas* 15: 5–25. Available at: http://casus.usal.es/pkp/index.php/MdE/article/view/1032/pdf_2 (accessed February 2019).

Pina-Sánchez, J., J. Koskinen, and I. Plewis. 2014. "Measurement Error in Retrospective Work Histories." *Survey Research Methods* 8: 43–55. Doi: https://doi.org/10.18148/srm/2014.v8i1.5144.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "Coda: Convergence Diagnosis and Output Analysis." *R News* 6: 7–11. Available at: http://oro.open.ac.uk/22547/ (accessed February 2019).

Poterba, J. and L. Summers. 1995. "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification." *Review of Economics and Statistics* 77: 207–216. Doi: https://doi.org/10.2307/2109860.

Poterba, J.M. and L.H. Summers. 1984. "Response Variation in the CPS: Caveats for the Unemployment Analyst." *Monthly Labor Review* 107: 37–43. Available at: https://stats.bls.gov/opub/mlr/1984/03/rpt1full.pdf (accessed February 2019).

Pyy-Martikainen, M. and U. Rendtel. 2009. "Measurement Errors in Retrospective Reports of Event Histories. A Validation Study with Finnish Register Data." *Survey Research Methods* 3(3): 139–155. Doi: https://doi.org/10.1002/sim.4780121806.

Richardson, S. and W.R. Gilks. 1993. "Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error." *Statistics in Medicine* 12(18): 1703–1722. Doi: https://doi.org/10.1002/sim.4780121806.

Rubin, D.B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American statistical Association* 91(434): 473–489. Doi: https://doi.org/10.2307/2291635.

Rubin, D.C. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, D.C. and A.D. Baddeley. 1989. "Telescoping is Not Time Compression: A Model." *Memory & Cognition* 17: 653–661. Doi: https://doi.org/10.3758/BF03202626.

Shiffrin, R.M. and J.R. Cook. 1978. "Short-term Forgetting of Item and Order Information." *Journal of Verbal Learning and Verbal Behavior* 17(2): 189–218. Doi: https://doi.org/10.1016/S0022-5371(78)90146-9.

Skinner, C. and K. Humphreys. 1999. "Weibull Regression for Lifetimes Measured with Error." *Lifetime Data Analysis* 5: 23–37. Doi: https://doi.org/10.1023/A:1009674915476.

Solga, H. 2001. "Longitudinal Survey and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison." *Quality and Quantity* 35: 291–309. Doi: https://doi.org/10.1023/A:1010387414959.

Veronesi, G., M.M. Ferrario, and L.E. Chambless. 2011. "Comparing Measurement Error Correction Methods for Rate-of-change Exposure Variables in Survival Analysis." *Statistical Methods in Medical Research* 22(6): 583–597. Doi: https://doi.org/10.1177/0962280210395742.

Wang, C.Y., L. Hsu, R.L. Feng, and Z.D. Prentice. 1997. "Regression Calibration in Failure Time Regression." *Biometrics* 53: 131–145. Doi: https://doi.org/10.2307/2533103.

Received July 2017
Revised July 2018
Accepted August 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 231–277, http://dx.doi.org/10.2478/JOS-2019-0011

Evidence-Based Monitoring of International Migration Flows in Europe

Frans Willekens[1]

In Europe, the monitoring and management of migration flows are high on the political agenda. Evidence-based monitoring calls for adequate data, which do not exist. The sources of data on international migration differ significantly between countries in Europe and the initiatives to improve data collection and produce comparable data, including new legislation, did not yield the expected outcome. Scientists have developed statistical models that combine quantitative and qualitative data from different sources to derive at estimates of migration flows that account for differences in definition, undercoverage, undercount and other measurement problems. Official statisticians are reluctant to substitute estimates for measurements. This article reviews the progress made over the last decades and the challenges that remain. It concludes with several recommendations for better international migration data/estimates. They range from improved cooperation between actors to innovation in data collection and modelling.

Key words: Europe; international migration statistics; migration flow modelling.

1. Introduction

The quality of international migration statistics in Europe has been an issue for decades. In the early 1970s, the Conference of European Statisticians (CES), a subsidiary of the United Nations Economic Commission for Europe (UNECE) and the United Nations Statistical Commission, noted serious shortcomings in statistics on immigration and emigration (Kelly 1987). The UN Economic Commission for Europe initiated a study comparing immigration and emigration statistics of member countries and found great discrepancies. While preparing demographic scenarios for Europe in preparation for the Conference on 'Human Resources in Europe at the Dawn of the 21st Century', Eurostat concluded that the existing data are inaccurate and not usable for population projections (Willekens 1994). Poulain (1991) had documented the inaccuracies. The migration flow

[1] Netherlands Interdisciplinary Demographic Institute, Lange Houtstraat 19 2511 CV Den Haag, Den Haag 2502 AR. The Netherlands. Email: Willekens@nidi.nl

Acknowledgments: Earlier versions of the paper were presented as an invited paper at the 2016 Conference of European Statistics Stakeholders (CESS), co-organised by the European Statistical Advisory Committee (ESAC), Budapest, 20–21 October 2016; an invited paper at the Eurostat conference "Towards more agile social statistics", session "Statistics on intra-EU mobility", Luxembourg, 28–30 November 2016, and as an invited keynote scientific speech at the 103rd DGINS Conference (Conference of the Directors General of the National Statistical Institutes), Budapest, 21 September 2017.
James Raymer (National University of Australia), Michel Poulain (University of Tallinn), Jakub Bijak (University of Southampton), Phil Rees (University of Leeds), Nathan Menton (UNECE), two reviewers and an Associate Editor of Journal of Official Statistics (JOS) provided extensive comments on earlier drafts. I am grateful to them for their comments and suggestions.

data could not be used and Eurostat used net migration estimates instead. That practice continues today (EUROPOP2015) (Lanzieri 2017a, 2017b). Net migration is obtained as a residual (population change minus natural change) without reference to data on migration. That approach allocates to migration the effect of several statistical adjustments made to balance the demographic accounting equation. Disregarding information on immigration and emigration has far-reaching implications, not only for demographic projections and the EU Economic Policy Committee's monitoring of the sustainability of public finances in EU Member States (which relies on Eurostat's population projections), but also for migration governance and the public debate on immigration.

The demand for accurate migration flow data increased ever since migration became a crucial issue for Europe and started to dominate policy and political agendas. The Amsterdam Treaty, adopted in 1997, requested the European Commission to develop uniform procedures for the management of international migration and for the production of community statistics, including migration statistics. The Treaty led to the establishment, in 2002, of the European Migration Network to promote the collection and dissemination of information on migration. In 2003, the European Commission and the European Parliament concluded that further progress towards improving migration statistics requires legislation. That resulted in new legislation in 2007, the *regulation on Community statistics on migration and international protection* (for further details on the history of Regulation (EC) No. 862/2007 of 11 July 2007, see Willekens and Raymer 2008). This legislation paved the way for statistical estimation methods, by allowing National Statistical Institutes to use estimation methods to produce the migration data to be submitted to Eurostat: "[a]s part of the statistics process, scientifically based and well documented statistical estimation methods may be used" (Article 9). Skaliotis and Thorogood (2007), both from Eurostat, discussed the challenges that migration posed to the European Statistical System. Regulation (EU) No. 1260/2013 of 20 November 2013 on the establishment of a common legal framework for the production of European demographic statistics in the Member States encouraged the use of scientifically based and well documented statistical estimation methods. The achievement of the objective of the Regulation, including the production of estimates, involves all Member States in an interactive way and effective coordination at the European level (Eurostat). The two Regulations and the Commission Implementing Regulation (EU) No. 2017/543 of 22 March 2017 on population and housing censuses also stress the need to harmonize concepts used in the production of statistics, in particular the concept of usual residence.

These developments and targeted funding by the European Commission, in particular Eurostat and the Directorate General for Research and Innovation, stimulated new research to improve the availability, reliability and comparability of migration data (for an overview of projects, see King and Lulle 2016; European Commission 2016a; Boswell 2016). In addition, NORFACE (New Opportunities for Research Funding Agency Cooperation in Europe) had a program to support migration research (Caarls 2016). The research resulted in an extensive assessment of data sources and the differences in the data produced, data collection practices, and activities undertaken at country and EU levels to overcome problems with migration data (Poulain et al. 2006; Kupiszewska and Nowok 2008; Kraler and Reichel 2010). In addition, improved statistical techniques were developed for estimating migration flows (e.g., Raymer and Willekens 2008; De Beer et al.

2010; Raymer et al. 2013; Abel 2013; Wiśniowski et al. 2016) and for forecasting migration in the presence of data deficiencies (Bijak 2011; Disney 2014). These studies did not yet resolve the inadequacies in migration statistics.

At the sixty-second plenary session of the Conference of European Statisticians in 2014, Lanzieri (2014a) of Eurostat reviewed research on European migration statistics and concluded that a wealth of methods is available to official statisticians for improving migration statistics, but that the potential remains under-exploited. Official statisticians are insufficiently aware of the methods that have been developed by researchers. Eurostat adds that the multiple methods studied and proposed may have created the impression that the research is not yet conclusive. Eurostat notes that the distinction between *statistics* and *estimates* hampers the implementation of research outcomes. Statistics represent the product of a compilation of records from primary data sources. Estimates represent the outcome of statistical models, possibly combining information from various sources. Official statisticians are reluctant to present estimates as official migration statistics, although the 2007 EC Regulation facilitated the use of statistical estimation methods to produce harmonized migration statistics. Eurostat calls for a strong and constant commitment to improve primary data sources and the derived statistics. Note that the compilation of records from primary data sources may also involve some estimation to overcome differences in definitions and measurements. By way of illustration, see the feasibility study by Statistics Netherlands on the production of migration data that satisfy the concept of usual residence specified in the European demographic regulations and the duration of stay criterion of 12 months (Statistics Netherlands 2016). Data from different sources, including the population register, are combined to produce consistent estimates.

In this article, I review recent research aimed at better data on international migration flows in Europe and argue that the most effective strategy to produce high-quality data on international migration for the monitoring and the management of migration is to create a *synthetic database*. A synthetic database combines quantitative and qualitative data from different sources. It contains the best possible estimates of the 'true' migration flows and indicators of how reliable the estimates are, given the different sources of uncertainty in the reported data. I argue that the development and maintenance of a synthetic database is a learning process, which implies that knowledge is updated in light of new evidence. The Bayesian model of learning combines data from different sources, while accounting for the uncertainties involved. These methods may ultimately be incorporated in the database leading to a *smart database*, which recognizes data types, suggests estimation methods and signals new trends and discontinuities in migration flows.

The structure of the article is as follows. In Section 2, I approach the development of a synthetic database as a learning process. Section 3 is a very brief overview of main data sources of international migration. The subject of Section 4 is the modelling of migration flows. The Poisson model is the dominant model of migration. It is a probability model that predicts count data and associates with each prediction a probability that the prediction coincides with observations. To estimate the parameters, different types of data, including expert opinions, may be used. Bayesian inference provides a formal framework for combining different data types. Sections 5 and 6 focus on different types of observation and the modelling of errors in observation. One observational issue is selected for an in-depth discussion: the duration threshold or duration criterion applied to define usual

residence and used in the definition and measurement of migration. Section 7 concludes the article.

2. Evidence Accumulation: A Learning Process

The reasons for the inadequacies of international migration statistics, identified by the CES in the 1970s (Kelly 1987), still exist today (Poulain et al. 2006; Lanzieri 2014a):

a. No common definition of immigration and emigration. Although Regulation (EC) No. 862/2007 requests member countries, whenever possible, to follow the United Nations recommendations on statistics of international migration (United Nations 1998), only a few countries adopt the UN definition of long-term and short-term migrant.
b. Coverage of migrants is often incomplete. In some countries, international migration statistics do not cover the entire resident population.
c. Undercount of migration continues to exist, in particular for emigration. By implication, return migrations are underreported too.

Data sources vary greatly between countries in Europe, even if some similarities exist. Some countries rely on the population census, other use surveys, and still other use administrative data, for example the population register, databases on residence and work permits, and border data. For a brief evaluation of administrative data other than population registers, see Poulain and Herm (2011). Population registers vary in accuracy because registration depends on self-reporting and therefore on the individual's willingness to report. Some countries introduced administrative adjustments to account for the undercount, while other did not. Countries also collaborate with other countries and share data on arrivals and departures to enhance consistency in international migration statistics. Mirror statistics, that is, statistics produced on the same subject by other countries, explain and reduce asymmetries in reported international migration statistics.

The power of official statistics depends on the trust that stakeholders have in the figures. To be trustworthy, statistics should be valid, accurate, precise and reliable. Measurements are valid if they measure what they are supposed to measure. They are accurate if they represent reality. They are precise if different measurements yield results that are close. Measurements are reliable if they produce the same results under varying conditions. To produce international migration statistics that meet these requirements, direct measurements are necessary, but not sufficient. Direct measurements (primary data) should be complemented by scientifically based and well documented statistical estimation methods that make optimal use of the observations and quantify distortions and their effects on the derived statistics. An effective strategy is to create a *synthetic database* combining data from different sources and to view the development and maintenance of the database as a *learning process*. Learning involves a knowledge structure, the search for new evidence and integration of evidence in the knowledge structure.

2.1. Synthetic Database

Governments collect data for many nonstatistical purposes, such as tax and labour market policies. Other public and private organisations also collect data for purposes of administration and management. Some scientists collect data, but even if they do not, they

may have useful knowledge about migration flows. All these data can be used for statistical purposes. The European Commission (2009) supports the use of data from multiple sources, including the private sector, to improve statistics. The integration of different data types into a single synthetic database poses a major challenge. Large differences in definition and measurement of migration do not justify the production of migration statistics from raw data only. The data need to be harmonized. A useful harmonisation strategy is to use a model of migration that can accommodate different data types, both quantitative and qualitative data. The purpose of the model is to produce the best possible *estimates* of the 'true' number of migrations (by migrant category). Quantitative data come mainly from primary data sources (see following section) but may include previous measurements or estimates of migration flows, for instance data from a population census organized several years ago. Qualitative data include knowledge about migration flows elicited from subject matter experts. Estimates of true flows are updated when new data become available. An advantage of a model of true migration flows is that it can be used to simulate different types of data, including new forms of data, and different measurement methods. Models can also be used to assess the impact of data types and measurement methods on the discrepancy between true versus reported migration flows. The models can subsequently be integrated in migration forecasting (Disney et al. 2015).

The need for a model that integrates data from different sources has been set out in Eurostat's vision for the production of statistics (European Commission 2009). In that vision, an integrated model is proposed, in which needs for statistics are identified and the European Statistical System (ESS) attempts to respond to these needs by drawing upon, and integrating, information from different administrative and survey data sources (Radermacher and Thorogood 2009; Kraszewska and Thorogood 2010). Obtaining migration estimates that meet the expectations of stakeholders calls for a concerted effort. It cannot be achieved only at the national level, but needs to involve Member States in an interactive way, which requires effective communication, collaboration, data sharing, and coordination at the intra-European level.

2.2. *Learning Process*

The combination of data from different sources and the updating of prior knowledge in light of new evidence are essentially learning processes. Insight produced by one data source changes when data are added from another source. Viewing the development and maintenance of a synthetic database on migration as a learning process implies a *cognitive approach to database development*. The cognitive approach is currently the dominant approach to machine learning and artificial intelligence (cognitive computing). It could also be a useful approach to database development. A formal method of learning that is particularly useful in this context is the Bayesian model of cognitive development, in short Bayesian learning. A fundamental premise is that processes such as migration involve many uncertainties; the outcome (e.g., whether an individual migrates in a given period or the number of migrations in a population during the same period) is inherently uncertain. To process information effectively and produce reliable statistics despite the uncertainties is a challenge. The uncertainties imply that an outcome can take on a range of possible values. If the outcome is a discrete variable, a probability can be associated with each

possible value. If the outcome is a continuous variable, a non-zero probability can be associated with an interval. The distribution of probabilities indicates which outcomes are more likely and which are less likely. The more we know about a process, the better we are able to identify possible outcomes and predict how likely they are. The Bayesian model of learning is a formal approach to updating existing (prior) knowledge or beliefs in light of new evidence. Fundamental features of the Bayesian approach are that (1) knowledge or beliefs on processes and their outcomes are represented as probability distributions, and (2) when new evidence becomes available, the prior beliefs are updated. The Bayesian method is a probabilistic method of scientific reasoning (Howson and Urbach 1989). The method has shown to be effective in a range of areas, including cognitive science and statistics.

Bayesian learning involves a formal description of how new information is assimilated in existing cognitive schemes, that is, of the mechanism of integrating data from different sources into a coherent structure. It facilitates interpretation of data and it can also be used to study the measurement bias in existing cognitive schemes. These insights contribute to the production of valid, accurate and reliable information on a subject or process from empirical observation and prior knowledge. That makes Bayesian learning particularly attractive for the estimation of international migration flows.

Bayesian learning is remarkably similar to Piaget's theory of learning, known as constructivism. The theory states that people learn by incorporating newly acquired information or experience in the knowledge they already possess (see e.g., Miller 1983 for a good introduction to Piaget's theory). Both learning theories insist on the importance of prior beliefs and knowledge for the interpretation of new information and the prediction of unknown outcomes (Tourmen 2016, 14). According to Piaget, children and other individuals build (causal) models of the world in order to interpret observations and experiences and to predict what will happen next. Knowledge is structured and stored in mental structures, known as cognitive *schemes*. Schemes are structured knowledge representations in our mind. They are mental models of reality. They represent the knowledge base an individual relies on to interpret observations and experiences and to make predictions, in short, to make sense of the world. They determine an individual's beliefs about the processes in his or her environment (world view) and how these processes are perceived. New experiences and evidence usually lead to updating the cognitive schemes. *Assimilation* is the incorporation of new experiences into an existing framework without altering that framework. As long as new observations and experiences are aligned with the internal representations of the world, they can be assimilated and the mental model is adequate for interpretation and prediction. If new evidence contradicts an individual's internal representation, the individual may (a) disregard the evidence (denial), (b) change his or her perception of the evidence to fit the internal representation, or (c) adjust the mental representation. Piaget refers to the adjustment of knowledge structures in the light of new observations or experiences as *accommodation*. The processes of assimilation and accommodation describe a learning mechanism. Learning is building and updating cognitive schemes, a process known as constructivism.

Piaget did not elaborate on how knowledge is stored in mental schemes. In the Bayesian method of learning, knowledge is stored as probabilities and probability distributions. Beliefs are subjective probabilities associated with given outcomes or events. Subjective

probabilities are updated in light of new evidence. The similarities between Piaget's theory of learning and the Bayesian method have recently attracted the interest of cognitive scientists (see e.g., Frank 2016; Tourmen 2016). Learning processes in humans and machines are increasingly being formalized as Bayesian probabilistic inference (e.g., Chater et al. 2006; Gopnik and Tenenbaum 2007; Perfors et al. 2011; Jacobs and Kruschke 2011; Gopnik and Bonawitz 2015).

3. Sources of Information on Migration

The main data sources for international migration are censuses, administrative records and sample surveys (for a general introduction, see for example, Bilsborrow et al. 1997; Cantisani et al. 2009; Bilsborrow 2016). At the world level, the population census is the main data source. The census reports, for members of the resident population, the current place of residence, that is, at the time of the census, and the place of birth. These data make it possible to distinguish between native- and foreign-born. The census may also solicit from respondents the place of residence one or five years prior to the census or the duration of residence and the previous place of residence. The United Nations, the Organisation for Economic Co-operation and Development (OECD) and the World Bank have invested in making these census data publicly available. The quality of data varies because not all countries adhere to the UN Recommendations for Population and Housing Censuses. Some features of the census limit the usefulness of the census as a source for up-to-date data on migration flows (Willekens et al. 2016). First, the census obtains information from the *resident* population. Hence immigrants are included, but emigrants are not. The number of emigrants from a country may be derived from censuses of destination countries (mirror data), provided the country of birth is reported (Dumont and Lemaitre 2005). Second, the age or year of migration cannot be derived from the date of birth. Hence, unless data are available on place of residence at some recent date prior to the census, the data are ill-suited for an analysis of migration trends and effects on migration of social, economic or political events and processes, and natural disasters. Third, return migrations and frequent migrations go unnoticed. Fourth, censuses come only every ten years in most countries. In Europe, the traditional census is being replaced by a register-based census. In a register-based census, the census is conducted on the basis of information in the registers, rather than through field enumeration. Information in registers may be complemented by data from other sources. Valente (2010) reviews census-taking in Europe.

Abel (2013) developed a method to estimate international migration flows from census data on place of current residence and place of birth. The estimates are counts of people that changed residence at least once during a period of fixed length prior to the census (see also Abel and Sander 2014; Abel 2016). Lanzieri (2014b) of Eurostat tested whether Abel's method can be used to overcome problems of quality and availability of migration data in Europe. The test showed that the method cannot provide a full coverage of migration flows within the EU-EFTA region, primarily due to lack of input data, but can estimate the flows of persons born in specific countries. Lanzieri also found that the method can profitably be applied using any breakdown of population stocks, such as by citizenship or educational attainment.

Administrative data are produced by organisations in connection with administrative procedures. People have to register their residence status and their address when they enter school, apply for a work permit, a driver's license or social security. They are required to report any change of address. Several countries keep a population register, an individualized data sheet (personal card) that includes a unique identification number, personal characteristics, and a continuous registration of a selection of life events. When newborn children and immigrants are registered, a data sheet is created. Deaths and emigrations result in de-registration, provided people notify the local authorities that maintain the register. The population register is used for a range of administrative purposes and, when kept up-to-date, is a tool to track individuals and retrieve data at the individual level. The population register may be linked to other administrative data, for example, business register, housing register, register of residence permits and working permits, to individual data collected by censuses and surveys, and to administrative data collected by private organisations. Although administrative data are not collected to monitor population change, a selection of administrative data is provided to statistical institutes to produce statistics. The timeliness of the updating of the population register and the accuracy of the information determine the quality of the derived statistics. For a discussion on the potential of population registers for migration statistics (and other demographic statistics), see Poulain and Herm (2013). In addition to the registration data mentioned, other registration data are useful for migration statistics, for example, register of visa recipients and asylum seekers.

Sample surveys provide relatively detailed data on a selection of individuals. The information is usually collected at one point in time only (cross-sectional survey). In some surveys, individuals are followed over time and information is recorded at regular intervals (panel surveys, follow-up studies). Although surveys may include information on current and previous places of residence, the sample size is usually too small to determine the level and direction of migration in a population. However, surveys may yield a wealth of information on respondents and that information may be used to determine who is likely to migrate and who is not, and why. Migration data are extracted from household surveys, labour force surveys (Wiśniowski 2017), and surveys on living conditions (see e.g., De Brauw and Carletto 2012). Several of these surveys include questions on place of birth and previous place(s) of residence. Some solicit information on household members living abroad. Recently, Bocquier (2016) assessed whether in developing countries, demographic surveys and demographic and health surveillance systems can be sources of migration data. In the area of gender statistics, it is common to collect data on gender in general social and economic surveys. Eurostat proposed a similar approach for migration (Knauth 2011) and in 2010 the European Statistical System Committee (ESSC) adopted a conceptual framework and work program for migration statistics mainstreaming and the development of migration statistics. Mainstreaming of the migration dimension in data collection has great potential, not only for the production of migration statistics but also for socio-economic policies and development cooperation.

Designated migration surveys exist too. Designated surveys yield better insight in (a) the who, why and how of migration, and (b) effective policies aimed at the management of flows (Willekens et al. 2016). They differ from migrant surveys, which focus on migrants. Examples of designated migration surveys include the International Passenger Survey

(IPS) in the United Kingdom, the Migration between Africa and Europe (MAFE) survey, and the Mediterranean Household International Migration Survey (MED-HIMS). The IPS is used to determine the number of immigrants and emigrants of the United Kingdom. It is the main source of international migration statistics in the United Kingdom. A selection of travelers is asked how long they intend to stay in the United Kingdom or away from the United Kingdom (ONS 2015). Intentions may change and the Office of National Statistics (ONS) estimates the number of 'switchers'. To predict the number of people who stay at least 12 months in the United Kingdom or abroad (long-term international migrant LTIM), the ONS computes for each respondent in the IPS, "a person's probability to switch their intentions based on their nationality and the average number of people who have switched their migration intentions in the previous three years." (ONS 2016, Annex 1).

The MAFE was organized in 2008 in three countries of Africa and six countries of Europe to gain insight in reasons for migration, the methods people use to enter Europe, and the impact of personal contacts on migration (Beauchemin 2018). MAFE survey data have been used to estimate rates and probabilities of emigration from countries of Africa to Europe, using extensions of statistical techniques of event history analysis that account for complex sample design (oversampling of migrant households) (Schoumaker and Beauchemin 2015; Willekens et al. 2017).

In a MEDSTAT (European Commission's statistical cooperation programme for the countries of North Africa and the Eastern Mediterranean) regional workshop in Wiesbaden in March 2008, participating countries called for the implementation of a household migration survey to overcome the lack of data on international migration for the Mediterranean (MED) region (MEDSTAT Committee for the Coordination of Statistical Activities 2011). The MED-HIMS (Households International Migration Surveys in the MED countries) questionnaire is designed to collect data on out-migration, return migration, forced migration, intention to migrate, circular migration, migration of highly-skilled persons, irregular migration, and other useful data on migration, migrants, and the effects of migration on households and communities. National statistical offices implement the surveys. The countries covered by MEDSTAT are: Algeria, Egypt, Israel, Jordan, Lebanon, Morocco, Syria and Tunisia, as well as the Palestinian Authority. So far (August 2017), the MED-HIMS survey has been implemented only in Jordan and Egypt (Eurostat 2017). For a description of the project in the context of other international migration surveys, see Bilsborrow (2016). Designated international migration surveys have common goals, use common methods and face similar challenges of sample design, questionnaire design, implementation, data processing, and analysis. To gain insight into migration flows and their root causes, scientists recently called for a World Migration Survey (Beauchemin 2013, 2014; Bilsborrow 2016; Willekens et al. 2016). The survey could build on the experiences gathered in the MAFE and MED-HIMS surveys and other multi-country international migration surveys, such as the Mexican Migration Project (MMP) of Princeton University and the Push-Pull Project, a joint venture of Eurostat and the Netherlands Interdisciplinary Demographic Institute (NIDI) (Schoorl et al. 2000; Van Dalen et al. 2005). The promises and challenges of survey-based comparative international migration research have been documented, and the experiences and lessons learned reviewed (Liu et al. 2016). A World Migration Survey would be a significant step toward an

understanding of why people leave their home country and what should be done to develop a sustainable system of global migration governance.

New technologies lead to new forms of data. Mobile phones and other internet-connected devices generate data on the geographic location of the object. Geolocation data constitute a new form of data, obtained from a variety of sources, such as Global Positioning System (GPS) signals, the physical addresses associated with Internet Protocol (IP) addresses, and RFID (Radio-Frequency Identification) tags attached to objects (e.g., passports or identity cards). Internet Protocol (IP) addresses have been used to map locations from where users sent e-mail or used social media within a given period. Twitter and Facebook data, and Yahoo! email accounts have been used to infer migration flows. Google search data have been used to infer migration intentions and preferred destinations. Recently, Fiorio et al. (2017) used Twitter data to estimate the relationship between short-term mobility and long-term migration. Gerland (2015) and Hughes et al. (2016) review estimations of migration flows from geolocation data. Although geolocators track the locations of online connections and not the addresses of users or owners, and IP addresses can be masked, geolocation data may complement traditional data sources, provided they are available on a regular basis, anonymous, and the selection bias and privacy issues can be resolved. The challenges of using geolocation data as a source of migration data are huge (Laczko and Rango 2014). Hughes et al. (2016, 29) conclude that "[n]ew and traditional data sources do not substitute for each other, they complement each other. . . . Combining data sources is key to producing an infrastructure that is robust to unanticipated changes in the use of technology. Building that infrastructure would be a gradual and incremental process where increasing data production and access, together with the development of methods, would sustain each other. We believe that Bayesian statistical models for migration count data hold the promise of addressing the issue of unifying traditional and emerging data sources." The view that the new forms of data, known as *big data*, may complement but not replace traditional data sources, is consistent with the vision of the European Statistical System (2015).

4. Modelling Migration

The oldest model of migration is the gravity model. It predicts migration flows from characteristics of place of origin and place of destination, and the distance between origin and destination. Characteristics include population size. Distance is usually physical distance, but can also be cultural distance. The gravity model is deterministic and lacks quantification of uncertainties in the measurement of migration. In the early 1980s, researchers reformulated the gravity model as a probability model, more particularly a Poisson regression model (see e.g., Flowerdew and Aitkin 1982; Willekens 1983). The advantages were that (i) the gravity model could easily be extended by including a range of predictors of migration, (ii) the theory of statistical inference could be used to estimate the parameters of the model, and (iii) the data generating process is specified (implicitly or explicitly). That process, which is assumed to generate observations on migration numbers, is a stochastic process, more particularly a Poisson process (Pinsky and Karlin 2011, chap. 5) (see further). The Poisson regression model is the most popular model of migration. It is usually written as a log-linear model, with the log of the number of migrants as the dependent variable. The log-linear model is a member

of the family of generalized linear models (GLM). For an introduction to the Poisson model and other probability models of migration, see, for example, Willekens (2008, 2016a). For applications of Poisson regression models in estimations of true unknown migration flows in Europe, see Abel (2010), Raymer et al. (2013) and Wiśniowski et al. (2013). Cohen et al. (2009) apply the Poisson regression model (presented as GLM) to estimate migration between selected countries and regions of the world.

The assumption that migration flows are outcomes of an underlying Poisson process is restrictive. The Poisson distribution is fully determined by a single parameter: the expected number of migrations during a given period, for example, a year. The variance of the Poisson-generated flows is equal to the expected value of the flows. If migration flows are small, as in international migration, the variance in the data is usually much larger than the variance implied by the Poisson process. To account for larger variance or overdispersion, the negative binomial distribution is often used (Davies and Guy 1987; Congdon 1993). Abel (2010) and Ravlik (2014) use the negative binomial regression model to predict international migration flows. The negative binomial emerges as a limiting case of a mixture of Poisson distributions where the mixing distribution of the Poisson parameter is a gamma distribution. The study of overdispersion in migration data could benefit from developments in other fields, such as biostatistics. Payne et al. (2017) review several methods for dealing with overdispersion and Chebon et al. (2017) list three factors that contribute to overdispersion in count data: (1) unobserved heterogeneity due to missing covariates, (2) correlation between observations (such as in longitudinal studies), and (3) the occurrence of many zeros (more than expected from the Poisson distribution). In mobility studies, the mover-stayer model is the earliest example of a mixture model that accounts for the unobserved differences between movers and stayers (see e.g., Goodman 1961). In Section 6 of this article, I discuss the mover-stayer model in the context of international migration. Correlation between observations may be associated with factors that generate spatial dependence and spatial structure. It leads to spatial autocorrelation (for a discussion, see Griffith and Haining, 2006). In the presence of many zeros (*zero-inflated data*), which is relatively common in migration tables, the zero-inflated Poisson (ZIP) model may be used. It consists of two components. The first is a binary regression model that predicts structural zeros. The proportion of structural zeros is a latent variable. The second is a Poisson model that predicts counts for the remaining observations. The mover-stayer model is essentially a ZIP model (Yiu et al. 2017).

Not all scientists quantify uncertainty (e.g., Poulain 1993; De Beer et al. 2010). Those who do quantify uncertainty, do not all specify a Poisson model or its extension, the negative binomial model. Bijak (2011, 96) explicitly deviates from the Poisson model in favor of a normal distribution. Brierley et al. (2008, 153) assume that observations on migration flows follow a log-normal distribution with, as expected value, the log of the true flow and a given variance reflecting undercounting and other sources of uncertainty (log of data are normally distributed around the true values with a common assumed variance). True flows are predicted by push and pull factors. Azose and Raftery (2015) and Azose et al. (2016) focus on net migration and do not refer to the underlying process generating the migration flows. They predict net migration from past net migrations.

Today, the common approach to the estimation of migration is to specify a model of flows and to determine the unknown parameter values that maximize the *probability that the*

model predicts the observed flow data. The number of migrations (by characteristics of persons migrating, by origin and destination, during a given period) is the dependent variable of the model. In the statistical literature, that data type is referred to as *count data* and the stochastic process generating the data is a *counting process*. A counting process is a stochastic process that counts the number of events as they occur. A model with parameter values that are not plausible is not likely to yield accurate predictions of migration flows. The most common method to determine the unknown parameter values is to maximize the likelihood function. The model of migration flows relates migration to (a) factors that (are assumed to) influence migration systematically, and (b) random factors. The effects of random factors are captured by specifying an appropriate stochastic process. For instance, if $N(t)$ is a random variable denoting the migration count in year t or during the period from 0 to t, then the sequence $\{N(t)\} = \{N(0), N(1), N(2), \ldots.\}$ is a counting process. Counting processes arise in different ways, for example, by counting the number of times a person migrates before a given age x, or by counting the number of persons who migrate in a given period. The migration flow model should be consistent with the postulated underlying stochastic process. The implication is that the mathematical structure of the model of migration is determined by the assumed underlying stochastic process.

Many statistical models are based on counting processes. The theory, which was developed by Aalen (1975) in his PhD thesis, is well-established (Andersen et al. 1993; Aalen et al. 2008). It emerged as the main statistical theory for the estimation of models of event occurrences (survival models), event sequences (event history models) and complete life histories (for a brief introduction and for applications see e.g., Willekens 2014). The Poisson process is the simplest and most widely used counting process. It has a single parameter, the expected value of the number of migrations in an observation period. The variance is equal to the expected value. If events occur randomly in continuous time and if the occurrences are independent of each other, then the counting process is a Poisson process. The parameter of the Poisson process may vary by age, sex, income, region of origin, region of destination, and other factors. The parameter may also vary in time. For each of these categories, the parameter may follow a probability distribution to reflect the unobserved heterogeneity in a population.

By way of illustration, consider a change of residence and disregard the restriction on duration of stay associated with the concept of usual residence. I refer to a change of residence without duration threshold as *relocation*. An individual may relocate multiple times during a period of observation. Hence, relocation is a repeatable event. Let $N(t)$ denote the number of relocations experienced by the individual during t years of observation, from onset at time 0 to time t. Assume that relocation is governed by a Poisson process. That implies that the count variable $N(t)$ is a Poisson random variable and the distribution of possible values of $N(t)$ is a Poisson distribution. Without loss of generality, we assume that people are identical with respect to their relocation behaviour, which implies that all have the same propensity to relocate. The likelihood of observing n relocations between 0 and t is given by the Poisson distribution:

$$\Pr\{N(t) = n|\lambda\} = \frac{\lambda^n}{n!} e^{[-\lambda]} \tag{1}$$

The parameter of the Poisson distribution (λ) is the expected number of relocations during the observation period ($\lambda = E[N(t)]$). The variance is also equal to λ: $\mathrm{Var}(N(t)) = \lambda$. The

value of λ is determined by maximizing the probability that model (1) predicts the observations (maximum likelihood method).

The *relocation rate* is the number of relocations per individual per year. It is the ratio of the observed total number of relocations by the study population during a given observation period (n) and the total duration of exposure (in years) by all individuals exposured to the risk of migration during that period (PY). The relocation rate is $\hat{\mu} = n/PY$, while: $\hat{\lambda} = \hat{\mu}PY = n$. Since relocation is a repeatable event, an individual remains at risk after a relocation, hence all people are at risk during the entire period irrespective of the numbers of relocations experienced. If people enter the population after the start of the observation period or leave the population before the end of the observation period, then the duration of exposure needs to be adjusted for late entry (left truncation) and departure (right censoring). The relocation rate μ is an occurrence-exposure rate. Note that $\lambda = \mu PY$. The likelihood of n events is proportional to $\mu^n e^{[-\mu PY]}$ since the exposure level PY is known. In Poisson regression models, PY is known as offset.

The estimation of the expected number of relocations during the observation period (λ) from the observed number of relocations illustrates the traditional approach to the prediction of migration flows. Frequently, relevant information about relocations and migrations is available from other sources and hence, not contained in the data. For instance, migration flow data may be available for some past year or period, for example, from a census. Subject matter experts may have relevant information that is not contained in the data, for example information on regulations introduced during the observation period that affect the registration of relocations and migrations or that cause a discontinuity in the relocation rate. Traditional models of migration often incorporate relevant prior information into the model. Algorithms to integrate historical data on migration in estimations of migration flows include the iterative proportional fitting (IPF) method, entropy maximisation and the EM (Expectation-Maximisation) algorithm (for an overview of these methods, see Willekens 1999).

To include prior information in the prediction of migration, most researchers today adopt the Bayesian approach to statistical inference. The approach postulates that some prior information is available on the unknowns (the true flows or the parameters of the Poisson model) and that the prior information comes as probability distributions of plausible values of the unknowns. The prior information can be objective, such as migration data of an earlier period, or subjective, such as expert opinions or beliefs. Fundamental features of the Bayesian approach are that (1) information and knowledge are represented as probability distributions of possible values, and (2) prior information on unknowns is updated in light of (new) observations. Prior information is expressed as a probability distribution. It implies an assumption that not only the expected value of a variable of interest is known, but that the distribution of possible values of the variable is known too. In traditional methods that use prior information (e.g., IPF), prior knowledge is represented as point estimates; the distribution is not considered. If the prior information is limited, a uniform distribution is appropriate because it assigns equal probabilities to all possible migration counts. This prior is said to be noninformative. When more evidence (data) becomes available, beliefs about the number of migrations are updated. The updates are captured in a posterior probability distribution. Updating

beliefs, opinions, knowledge or predictions in light of new evidence is essentially a learning mechanism.

To combine data and prior information on the unknowns, Bayes' theorem is applied (for an excellent and accessible introduction, see Bijak and Bryant 2016; for a textbook see Congdon 2001):

$$p(unknowns|data) \propto \frac{p(data|unknowns)\, p(unknowns)}{p(data)} \tag{2}$$

The p's denote probability distributions, that is, probabilities or probability density functions. The term $p(data|unknowns)$ is the probability that a migration model with unknown parameters predicts the data, that is, the observed migration flows. It is the likelihood function described above. The term $p(data)$ is the probability of observing the data. If the data are obtained by sampling a population, then it is the probability of obtaining that particular sample. The term $p(data)$ is fixed for any given data set and plays a minor role in most applications. It is often omitted. The term $p(unknowns)$ is the prior probability distribution. It represents empirical evidence (objective) or beliefs (subjective) about the values of the parameters of the model prior to data collection. In case of a noninformative prior, the posterior distribution $p(unknowns|data)$ is determined by the likelihood and the Bayesian method produces results that are similar to the traditional method. 'Unknowns' can be replaced by 'model' or 'hypothesis' in which case the prior is the probability that we select a model or formulate a hypothesis, given the data and prior information. The posterior probability distribution is used to determine a credible range of values of the unknowns (credible set), analogous to the confidence interval in classical (frequentist) statistics. A 95% credible interval has a 95% probability of containing the true value (Bijak and Bryant 2016, 3; Congdon 2001, 6). The range of values reflects the subjective prior beliefs, not only uncertainties in the data.

To illustrate the Bayesian approach to the estimation of migration, consider the likelihood function (1). Assume we have subjective prior information on λ that we want to use in the estimation procedure. We believe that λ is nonnegative and the possible values follow an exponential distribution (from 0 to ∞) with parameter ξ equal to 1, hence $p(\lambda|\xi) = \xi e^{[-\xi\lambda]}$ with $\xi = 1$, hence $p(\lambda) = e^{[-\lambda]}$. Given the distribution, the expected value of λ (the expected number of relocations during the period of observation) is $1/\xi = 1$, which may be very different from the number of relocations observed in the sample population. Suppose that, prior to data collection, we expect 1 relocation during the period of observation. The posterior distribution of the number of relocations is

$$\Pr\{\lambda|N(t) = n\} = \frac{\frac{\lambda^n}{n!} e^{[-\lambda]} e^{[-\lambda]}}{\int_0^\infty \frac{\lambda^n}{n!} e^{[-\lambda]} e^{[-\lambda]} d\lambda} = 2^{n+1} \frac{\lambda^n}{n!} e^{[-2\lambda]} \tag{3}$$

which is the probability density function of the gamma distribution with shape parameter $n + 1$ and scale parameter 1/2. The inverse of the scale parameter is known as rate parameter, in particular in the context of the Poisson process. Let b denote the scale parameter and c the shape parameter. A common specification of the gamma distribution

is (Evans et al. 2000, 98):

$$\Pr\{\lambda|b,c\} = \frac{(\lambda/b)^{c-1}}{b\Gamma(c)}\, e^{[-\lambda/b]} \tag{4}$$

where $\Gamma(c)$ is the gamma function. Since c is a positive integer, $\Gamma(c) = (c-1)!$ with ! denoting factorial of $c - 1$. The expected value of λ is $E[\lambda] = bc$, hence the expected posterior value of λ is $(n + 1)/2$, which is the mean of (a) the prior guess of the number of relocations during the observation interval and (b) the observed number. If we believe or assume that one individual relocates during a given period, but we observe 150 relocations, then the expected posterior number of relocations is 75.5. The central, for example, 95% interval of a gamma distributed random variable is obtained analytically from the relationship between the gamma distribution and the Poisson distribution (for a recent discussion, see Fay and Kim 2017). It can also be obtained by sampling from the gamma distribution.

The exponential distribution is a special case of the gamma distribution. It is a gamma distribution with $c = 1$ and b the inverse of the rate, the parameter of the exponential distribution. If the prior is a gamma distribution, the posterior is a gamma distribution too. The posterior and prior distributions are conjugate distributions and the posterior has a closed-form expression. For instance, if we assume a gamma prior for μ, then the posterior density for μ will be a gamma too. If the prior is $G(a,b)$, then the posterior is $G(a + n,b + PY)$. (Congdon 2001, 35).

Except for simple cases such as the one presented here, the mathematical form of the posterior probability distribution is not known and the parameter(s) cannot be obtained analytically. The solution is to explore the (joint) posterior distribution of the unknown(s) by walking around on that distribution (surface), take samples and determine how likely the samples are given the migration model, the prior distribution of the unknowns and the data. The walk is a random walk modified by an acceptance rule. The rule states that a proposed move from the current location to a new location is accepted if that move contributes to finding the target posterior distribution. Once the target distribution is found, samples are taken to determine the unknowns. The samples are not independent. The current location determines the new sample. That is operationalized by considering each possible location as a state in a state space. The sequence of states is a Markov chain. The transition probabilities of the Markov chain are the probabilities of accepting moves, which are determined by the acceptance rule. The Markov chain that results has the target posterior distribution as its equilibrium distribution. This method is the Markov Chain Monte Carlo (MCMC) method (see e.g., Congdon 2001, 466; Brooks et al. 2011; Bijak 2011, 32ff).

The MCMC method was developed in the 1940s by Metropolis (Metropolis et al. 1953) and extended by Hastings (1970). German and German (1984) introduced Gibbs sampling into the arena of statistics. The idea of Gibbs sampling is to simulate from conditional distributions to produce samples from a joint distribution. Software for MCMC simulations is abundantly available. In a chapter contributed to Bijak (2011), Wiśniowski reviews available software for Bayesian analysis. Popular software includes WinBUGS, OpenBUGS and JAGS. The BUGS platform was developed by the BUGS (Bayesian inference Using Gibbs Sampling) software project (www.mrc-bsu.cam.ac.uk/software/bugs/).

Increasingly, the R software environment is used for handling Bayesian computations. A good starting point is the text by Robert and Casella (2010) and the CRAN Task View "Bayesian inference" (https://CRAN.R-project.org/view=Bayesian).

5. Modelling Measurement Errors with Input from Subject-Matter Experts

Four key data problems emerge in the measurement of international migration (Raymer et al. 2013; Disney 2014; Disney et al. 2015): (1) the definition of migration, (2) population coverage (some population groups are omitted), (3) underreporting of migration, and (4) concerns about accuracy of the measurements. The prediction or nowcasting of the true migration flows in a given observation period by country of origin and country of destination is complicated by the mentioned measurement problems. The first is the definition of migration. Two broad data types are distinguished to define migration (Courgeau 1973): event data and status data. Event data measure event occurrences, for example, migrations. Status data measure personal attributes, for example, place of residence. By comparing places of residence at two points in time, the occurrence of a migration can be inferred. To distinguish these indirect measurements of migration from event data, they are referred to as transition data (see e.g., Willekens 2016a). A major source of event data is the population register. The population census and labour force surveys are major sources of transition data. Some important data problems in the measurement of migration can be reduced to these two data types (Willekens 1999; Poulain 2008). Event data and status data on migration are not really comparable, but they can be made comparable by modelling the 'true' migration process underlying both event data and status data.

The definition of migration is two-dimensional. It includes a spatial dimension and a temporal dimension. The spatial dimension defines the areal units (places of residence) considered in the measurement of migration. In international migration, it is a country. The temporal dimension adds a duration criterion to the definition of residence and change of residence. In 1998, the United Nations introduced the concepts of long-term and short-term migrant and adopted the definition of resident (as opposed to visitor) included in the 1994 United Nations recommendations on tourism statistics (United Nations 1998). A long-term migrant is a person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes his or her new country of usual residence. A short-term migrant is a person who moves to a country other than that of his/her usual residence for a period of at least three months but less than 12 months. The concept of short-term migrant also depends on the reason for migration. Moves that are for purposes of recreation, holiday, visits to friends and relatives, business, medical treatment or religious pilgrimage are excluded. Many countries do not follow the UN definition, but use different duration thresholds. Some, for example Germany, have no threshold and consider all changes in usual residence as migrations irrespective of intended or effective duration of stay. Other countries, for example Poland, register a change in usual residence if and only if a person indicates that the change is permanent. For a list of EU and EFTA countries and the duration thresholds they consider in measuring migration, see Cantisani and Poulain (2006); UNECE (2012) and Raymer et al. (2013, 803). The UNECE Task Force on

Analysis of International Migration Estimates Using Different Length of Stay Definitions, set up in 2008 to explore the different definitions in use, found five different ways countries in Europe measure duration of stay for an immigrant and duration of absence for an emigrant (UNECE 2012). Regulation (EC) No. 862/2007 on migration adopted the definition of long-term migrant recommended by the United Nations. Regulation (EU) No. 1260/2013 of 20 November 2013 calls on the Member States to carry out feasibility studies to determine whether the country can comply with the UN (and Eurostat) definition of usual residence (by 31 December 2016). It is worth noting that the International Organisation of Migration (IOM) does not adhere to the definition of migration proposed by the United Nations and omits duration of stay criteria (IOM 2011 and IOM website https://www.iom.int/key-migration-terms).

The definition of migration is complicated by differences in definition of *residence*. Countries that use a de jure enumeration of individuals register the usual place of residence, while countries that use a de facto enumeration record the actual place of residence. The concept of residence is increasingly becoming a fluid concept, one that means different things to different people. Some people, known as *transnationals*, have multiple residences in different parts of the world and identify with multiple communities. In other words, they have multi-sited individual and social lives (IOM 2010b; Bilgili 2014). Transnationalism is a key factor in contemporary migration management (IOM 2010a, 2010b). The definition of migration is also complicated by the concept of legal residence. A person who changes usual residence with an intention to stay at least 12 months in another country is not recorded as an immigrant unless the person is allowed to reside within the country of destination (and can show a document as proof of residency, such as a residence permit). Transnationalism and the concept of residency pose challenges for the definition of usual residence and the measurement of international migration. These challenges have their roots in the concept of sovereign nation state, introduced in the Peace Treaty of Westphalia (Germany) of 1648 as part of the new system of political order in Europe and upheld in the UN Charter. That treaty offers the legal basis to control national borders and regulate international migration (Betts 2011).

Measurement issues are also related to main method of data collection. In general, a population register yields better data on immigrations and emigrations during a given calendar year than surveys or other means of data collection. Censuses generally provide accurate data on immigrants but not on emigrants. A population register and a census differ in the residence concept used. A register considers the administrative place of residence, while the census uses the actual or usual place of residence. Countries with a population register differ in quality of the migration data. The quality is considered better in Nordic countries, which exchange individual data on international migration. The five Nordic countries record migration between the countries on a special form, *the Inter-Nordic Migration Certificate*, and pass individual data on new arrivals to the population register of the country of origin. To improve its migration statistics, Romania started to exchange aggregate data with Italy and Spain, two of the main destinations of Romanian emigrants (Pisică 2016). Two sources of error complicate the measurement of migration further: under-registration (undercount) and undercoverage. Undercount occurs when not all migrations are recorded. If immigration and emigration depend on self-reporting, the willingness to report varies, and the undercount can be substantial. Major sources of

under-registration of immigration are people who overstay their tourist visa or residence permit, and undocumented border crossing. Under-registration of emigration is caused by people leaving the country without notice. A consequence of under-registration of emigration is that return-migrations are under-registered. Some countries, for example the Netherlands, correct emigration statistics by including unreported emigration of foreigners if the administration reveals that residents are missing and likely moved abroad. Undercoverage occurs when some categories of the population are excluded from the measurement of migration. For instance, asylum seekers are usually excluded because they are not admitted yet to reside legally in the country, although they intend to stay at least 12 months. Countries differ in ways they record cross-border migration of nationals and foreigners (UNECE 2012). For instance, Romania's immigration data include foreigners only, while emigration data include nationals only (Romanian Institute for Research on National Minorities 2014). Because of differences in definition and measurement, a migration between two countries may be recorded in one country but not in the other. As a consequence, sending countries and receiving countries report different migration counts.

These measurement problems have been known for a long time and attempts to do something about it have a long history. The Conference of European Statisticians (CES) identified the problem as early as 1970. In 1971, the CES organized the United Nations European Seminar on Demographic Statistics in Ankara and Istanbul in cooperation with the United Nations Office for Technical Cooperation and the Government of Turkey (Kelly 1987). Participants noted that there were serious shortcomings in the statistics of immigration and emigration available for UNECE countries in that they differed considerably in scope, coverage, definitions, classifications, and content and that in most instances, they did not meet the requirements of population analysis research. They concluded that the improvement and harmonisation of statistics on international migration was an urgent task. They also recommended organizing an exchange of statistics on international migration among ECE countries. The CES followed the recommendation and the improvement of migration statistics was included in the 1972 work programme. In its 1974 meeting, the CES pointed out that the quality of immigration statistics is generally much better than that of emigration statistics and proposed that a meeting of interested countries be held to discuss arrangements for bilateral exchanges of data on migration between pairs of ECE countries. That meeting was organized in 1975. In preparation of that meeting, the ECE secretariat collected immigration and emigration statistics for 1972 and arranged the data in two origin-destination matrices, one based on immigration data and another based on emigration data. The bilateral flow data revealed serious asymmetries in the migration data compiled by the countries in the ECE region. Reported emigrations from country A to country B did not match the reported immigrations to country B from country A. In 1980, the Council of Europe collected similar data from the 21 Member States of the Council of Europe and found the same anomalies. The matrices were attached to the 1981 annual report on the demographic situation in Europe. Issues such as coverage and duration threshold were already discussed at that time. To date, these issues have not been resolved satisfactory, although considerable progress has been made. Actions called for in 1974, for example the exchange of statistics on international migration between countries, are still being called for today (Skaliotis and

Thorogood 2007; Radermacher and Thorogood 2009; Raymer 2012; Willekens et al. 2016), although an example of good practice exists; namely, the exchange system in the Nordic countries. Insights in the types of migration data being collected increased significantly (Poulain et al. 2006; Kraler and Reichel 2010) and methods for the reconciliation of national statistics have been developed. These methods are the subject of the remainder of this article.

Until concepts and definitions of residence and migration are refined and innovations in data collection methods and procedures reduce measurement errors and increase the comparability of data, methods are needed to infer trustworthy and comparable migration statistics from data provided by the different countries of Europe. Essentially, two methods have been developed to reconcile national statistics on international migration in Europe. Both start from the bilateral migration flow data compiled by countries of origin and countries of destination, and adjust the reported migration data to obtain a unique, complete and internally consistent matrix of migration flows between the countries of Europe. The first method adjusts the reported migration without explicit reference to the sources of error in the measurement of migration. The method was proposed by Poulain and Wattelar (1983) and improved by Poulain (1993, 1999). It was adopted as a point of departure in the Eurostat-funded project MIMOSA (Migration Modelling for Statistical Analyses 2007–2009), which resulted in several publications listed below. The method considers uncertainties in the data and experts help inform the estimation procedure by their judgments on the magnitude of the uncertainties that result from measurement problems. Migration flows to and from countries with good international migration data are given priority over migration flows between countries with serious data limitations and hence a larger uncertainty in migration flows. The second method pays more attention to the measurement process and specifies a *measurement model* that relates the quality of migration estimates to the main sources of measurement error: differences in definition, coverage, undercount and accuracy in migration measurement. Experts are interviewed and their judgments on the relative significance of the different reasons in explaining the incomparability of data are incorporated in the model. The measurement model is combined with a *migration model* that aims at predicting true migration flows (latent, not observed) from knowledge of the determinants of migration. The Bayesian approach is used to combine the different data types. The method was proposed by Raymer et al. (2013) in the context of the NORFACE-funded project IMEM (Integrated Modelling of European Migration 2009–2012). In the remainder of this section, I review the two methods.

5.1. The Poulain Approach with Extensions

Poulain and Wattelar (1983) proposed a method to reconcile immigration and emigration statistics. They distinguish three types of countries with different levels of data availability: (1) countries with immigration and emigration data, (2) countries with immigration or emigration data, and (3) countries without data on international migration. For each country, two correction factors are defined, one for immigration data and one for emigration data. Let I_{ij} denote the immigrants in j originating from i, reported by receiving country j, and E_{ij} the emigrants from i with destination j, reported by sending country i.

The correction factors correct the flow from i to j such that

$$\alpha_j I_{ij} = \beta_i E_{ij} \tag{5}$$

where α_j is a correction factor associated with the immigration data of country j and β_i is a correction factor associated with emigration data from i. If C is the number of countries, then there are $C(C-1)$ equations and $2C$ unknowns. The system of equations is overdetermined, that is, there are more equations than unknowns. To obtain an approximate solution, the Euclidean distance measure $\sum_{i,j}(\alpha_j I_{ij} - b_i E_{ij})^2$ is minimized subject to the constraint that the total sum of estimated migration flows is equal to the total of the observed immigrations:

$$\sum_{ij} S_{ij} = \sum_{ij} I_{ij} \tag{6}$$

where $S_{ij} = 0.5[\hat{\alpha}_{ij} I_{ij} + \hat{\beta}_{ij} E_{ij}]$

A two-step procedure is used to improve the quality of the estimates. In a first step, five countries with complete and relatively good data are selected and the correction factors are determined. The correction factors are fixed up to a constant. To remove the constant, one correction factor is set to a given value. The authors fix the correction factor of immigration data of Denmark to unity. A correction factor equal to one preserves the reported immigration data. In a second step, the correction factors determined in the first step are fixed and those for the other countries are determined. The procedure results in two migration flow matrices, one with corrected immigration data and the other with corrected emigration data. The two matrices are close, but not equal. Unique values of migration flows are obtained by averaging the corrected immigration flow and the corrected emigration flow. Poulain (1993) repeats the procedure, but considers two groups of countries. The first group consist of the Nordic countries with good data. The second group consists of the other countries. The procedure consists of three steps. First, the correction factors are estimated for the Nordic countries. In a second step, the correction factors for flows between the Nordic countries and the other countries are estimated. These factors are used in a third step to estimate the remaining migration flows. Poulain (1999) divides countries into three groups depending on the reliability of migration data. The procedure is similar to Poulain (1993). The approach implies that the estimates of the migration flows between countries with good data are not influenced by data of less quality.

Van der Erf and van der Gaag (2007) adopt the method developed by Poulain. They start with the Nordic countries and add countries successively based on the perceived reliability of their migration data. The sequence of countries introduced in the iterative estimation procedure is determined by experts. Expert judgments are also used to adjust correction factors if appropriate.

Poulain and Dal (2008) apply the procedure to estimate migration flows between 28 countries of Europe: 13 countries with consistent migration data (called referee countries) and 15 other countries. The correction factor for immigrations registered in Sweden is set equal to one because Sweden uses the UN definition of migration (12 months criteria) and is considered to record immigration accurately. They change the function to be minimized to $\sum_{i,j}(\alpha_j I_{ij} - b_i E_{ij})^2/(I_{ij} + E_{ij})$ and maintain a single constraint that the total averaged estimated flow is equal to the total immigration. The denominator removes a limitation of

the least square method, namely that large flows receive considerably more weight than small flows, which means that flows from large countries have a strong influence on the estimates. A limitation of that new distance function is that small flows receive much more weight than large flows. The problem is well-known in migration research and is resolved by considering multiple distance functions (Willekens et al. 1981; Abel 2010).

De Beer et al. (2010) adapted the constrained optimisation procedure to assure that the marginal totals of the corrected I and E matrices are equal:

$$\sum_j \hat{\alpha}_j I_{ij} = \hat{\beta}_i \sum_j E_{ij} \tag{7}$$

and

$$\hat{\alpha}_j \sum_i I_{ij} = \sum_i \hat{\beta}_i E_{ij} \tag{8}$$

for all i and j. Equations (7) and (8) can be written as a homogeneous system of $2C$ linear equations with $2C$ unknowns. The correction factors are unique up to a constant. The correction factor of reported immigration data for Sweden is set equal to one, for the reason given by Poulain and Dal. Since $\hat{\alpha}_{j=Sweden}$ is fixed, the system of equations becomes a system of nonhomogeneous equations of the form $Ax = B$, which has $2C$ equations and $2C - 1$ unknowns. The solution is of the form $x = A^g B$, where A^g is the generalized inverse of A. That solution is identical to the one obtained by minimizing $\sum_{i,j}(\alpha_j I_{ij} - b_i E_{ij})^2$ subject to constraints (7) and (8).

Missing data constitute a separate problem. Some authors omitted countries with missing data. Poulain used the correction factors obtained from countries with data to estimate migration flows for countries without data. Abel (2010) estimated the missing flows using a regression model, fitted to the harmonized international migration flow data. The predictors are characteristics (covariates) of sending and receiving countries, and characteristics of links between the countries (distance, contiguity, trade, and so on). The model is a spatial interaction model, an established type of model for estimating migration flows. The idea to introduce a migration model that relates the harmonized data to covariates was important and was adopted by others (e.g., Raymer et al. 2013). The parameters of the model are estimated taking into account the incompleteness of the observed (harmonized) data. Abel applies the EM (expectation-maximisation) algorithm, which is a maximum-likelihood technique that uses the migration model to predict missing data, initially with preliminary values of the parameters, and uses the predictions to improve the parameter estimates.

5.2. The Raymer et al. (IMEM) Approach

Raymer et al. (2013) use a migration model to predict true migration flows and a measurement model to quantify differences between observations and true flows. In their study, true migration flows are long-term migrations, that is, relocations for at least 12 months. A true flow can be defined in different ways, like events can be defined differently, as long as the definition is unambiguous and unique. The measurement model captures effects of the measurement problems mentioned above. The authors initially

assume that migrations are generated by a Poisson process, but they assume that the expected values of migration counts are normally distributed. That approach allows for larger variability than the Poisson distribution, that is, for overdispersion. The dependent variable of the migration model is $\log(\lambda)$, where λ is the true number of migrations (UN definition) in a given year. λ is origin-destination specific and is different for each calendar year. The true number of migrations in a given year is predicted by a set of covariates. A time-invariant normally distributed random factor (random effect) is introduced to smooth flows across time. The factor induces residual correlation between the same flows at different points in time. Variation in the random factor is restricted to induce a residual correlation between flows in opposite directions. If a flow is larger than predicted by the model, the flow in the opposite direction is also expected to be larger. The parameters of the model were estimated using the Bayesian method. Weakly informative prior distributions were used (normal distributions and gamma distributions with parameters fixed by the authors or drawn from probability distributions). This implies that the predictions of migration are driven mainly by the covariates and that the influence of prior information is limited. The prior distributions are selected for computational convenience only.

To convert the reported data to comply with the definition of migration used in the true data (UN definition), a measurement error model is introduced. The covariates are assumed to be measured correctly, but migration is not measured correctly for reasons listed above. Reported migration data, that is, the observations, are initially assumed to be generated by a Poisson process and to follow a Poisson distribution with parameter $\lambda*$ (by country of origin, country of destination and calendar year). The parameter $\lambda*$ differs from the parameters of the model of the true migration flows (λ) because of measurement errors. Immigrations and emigrations are modeled separately to account for the asymmetry in bilateral migration flow matrices, that is, substantial differences between immigration data from receiving countries and emigration data from sending countries. Let i denote the sending country and j the receiving country. Let λ_{ij} denote the true migration flow from i to j, Z_{ij}^R the flow from i to j reported in the receiving country j (immigration data), and Z_{ij}^S the flow from i to j reported in the sending country i (emigration data), λ_{ij}^{*R} the expected number of migrations from i to j recorded in j, and λ_{ij}^{*S} the expected number of migrations from i to j recorded in i. The expected number of migrations from i to j, observed in j, is proportional to the true flow:

$$\lambda_{ij}^{*R} = \lambda_{ij}\, a_{ij}^R \tag{9}$$

The proportionality factor a_{ij}^R depends on the factors that distort the measurement of immigrations in j. Equation (9) may be written differently, as $true\ flow = factor \cdot data$, with $factor = 1/a_{ij}^R$. The expected number of migrations from i to j, observed in i, is also proportional to the true flow:

$$\lambda_{ij}^{*S} = \lambda_{ij}\, a_{ij}^S \tag{10}$$

The proportionality factor a_{ij}^S is a function of the factors that distort the measurements of migration in country i: duration threshold, undercount, coverage, and country-specific level of accuracy of the data collection system.

a. Duration threshold

If the duration threshold is identical to the threshold used in measuring the true flow (12 months in the Raymer et al. study), then the threshold effect on the distortion is 1. If the duration threshold is less than the threshold used for the true data, true migration is overestimated and a_{ij} is larger than 1. If the duration threshold exceeds the one used in the true data, true migration is underestimated and a_{ij} is less than 1. Five duration threshold parameters are considered, one each for duration 0 (no duration threshold), three months, six months, 12 months and permanent. The duration threshold of 12 months is the reference category.

b. Undercount

The undercount effect is large if the undercount is large and small if undercount is low. IMEM considers two levels of undercount: low and high.

c. Coverage

The coverage effect captures country-specific deficiencies in measuring migration not reflected in the undercount. IMEM considers two types of coverage: standard and excellent. The coverage effect for a country is a normally distributed random variable, with mean and variance functions of the coverage assumed for that country (standard or excellent). To ensure that the random effect is between 0 and 1, a logistic transformation is used. Let k_i denote the normally distributed coverage effect $(-\infty,+\infty)$ for country i and let p_i denote the associated random effect between 0 and 1. Then $k_i = logit(p_i)$ and $p_i = 1 + e^{(-k_i)}$. For migration to and from the rest of the world, Raymer et al. assume perfect coverage. They justify that assumption by the more rigorous registration requirements for migrants originating from or departing to countries outside of the EU/EFTA region. That assumption and the justification are far from realistic.

d. Accuracy of the data collection system

A country-specific term is added to capture differences in accuracy of the data collection system, irrespective of duration threshold, undercount and coverage. IMEM distinguishes three types of data collection systems for migration: (1) registers in the Nordic countries, (2) other good register-based systems, and (3) less reliable register-based or survey systems.

Raymer et al. elicited opinions of experts on migration statistics to quantify the factors that distort the measurement of migration and to derive the prior for the estimation. Information was obtained from eleven experts using the Delphi method. Experts were invited to provide, for each distortion factor, (a) a range of values for the distortion, and (b) an indication of how certain they were about that range. The method is described in detail by Wiśniowski et al. (2013). For instance, consider the duration threshold. Experts were asked by how many percent they expect the level of migration with the six-month criterion to be higher than with the 12-month criterion, which is used to measure the true migration flow. They should not give a single percentage, but rather a range of percentages, for example, between 15 and 30%. The lower bound is $P_1^{(6)}$ and the upper bound is $P_2^{(6)}$. Hence the overcount factor ranges from $dur_1^{(6)} = 1 + P_1^{(6)}$ to $dur_2^{(6)} = 1 + P_2^{(6)}$. The beliefs of

experts need to be translated into a probability distribution. The authors considered probability density distributions for which the parameters could easily be calculated. To that end, an auxiliary variable d was introduced:

$$\begin{aligned} d^{(6)} &= \ln[dur^{(6)}] \\ d^{(3)} &= \ln[dur^{(3)}] - d^{(6)} \\ d^{(0)} &= \ln[dur^{(0)}] - d^{(3)} - d^{(6)} \\ d^{(p)} &= -\ln[dur^{(p)}] \end{aligned} \tag{11}$$

where p denotes 'permanent'. The expert-specific probability density of the auxiliary variable $d^{(s)}$, with $s = \{0, 3, 6, p\}$, is assumed to follow a log-normal probability density distribution. The mean and the standard deviation are estimated from the values of d, derived from the ranges of percentages given by the experts, weighted by elicited certainty levels. Wiśniowski et al. (2013, 598) show the equations. The individual densities were used to produce a mixed probability density distribution.

Raymer et al. (2013, 806) state that the median of the true flow (12-month duration threshold) is 81% of the median of the flow measured with the six-month duration criterion. The median of the true flow would be 51% of the median of the flow estimated with no time limit (overestimation 96%) and the median of the true flow would be 61% of the median of the flow measured with the three-month criterion. The median of the true flow would be 1.64 times the median of the flow estimated with the 'permanent' criterion.

A similar procedure was followed for the undercount and the coverage. For the undercount, the beta density was selected. The individual densities were used to produce a mixed density. The mean undercount of immigration was 72% with a standard deviation of 18%. The mean undercount of emigration was 56% with a standard deviation of 22% (Wiśniowski et al. 2013, 595). The large standard deviation and the flat shape of the distribution of the mixture densities reflects the heterogeneity of expert judgements about the undercount. Raymer et al. (2013) give further results. Experts believe that in countries with low undercount, 88% of the immigration and 73% of the emigration are reported. They believe that in countries with high undercount, 68% of the immigration and 45% of the emigration are reported. The lack of consensus among experts was an interesting finding. Wiśniowski et al. (2013) attribute it to different experiences of the experts with migration statistics. The experts' beliefs may have been based on the data collection systems they know best. A consequence of the lack of consensus among experts is that the probability distribution, if used as a prior in Bayesian estimates of immigration and emigration, is weakly informative, that is, not much different from a uniform distribution that attaches equal probabilities to all possible values. Wiśniowski et al. (2013, 603) conclude that the expert-based prior densities led to very wide posterior distributions of estimated migration flows. Expert-based prior densities do not produce estimates of migration that are substantially different from estimates based on noninformative or weakly informative priors. The authors list four lessons learned from the elicitation of expert opinions:

1. The form of the prior probability density distribution and the distinction between categories of countries matter.

2. The wording of questions posed to experts is important. Different formulations should be tested. Recently, Hanea et al. (2016) proposed the IDEA protocol as a method for removing linguistic uncertainties in eliciting expert opinions.
3. Certainty levels are easily misinterpreted. If an expert expects that, in a country, the undercount of immigration is between 20 and 35%, and the certainty level is 70%, then 30% of the immigrations are distributed outside of the range indicated by the expert. Several experts misinterpreted that mechanism.
4. One should be careful in selecting experts. Some invited experts were not convinced that subjective probabilities are useful information for the estimation of migration flows.

The authors do not question the usefulness of expert judgements to complement migration flow data, but propose a more thorough assessment of the empirical knowledge experts have and how they translate knowledge into subjective estimates of migration flows. In some cases, expert opinions may be replaced by models. In the next section, I discuss and illustrate the use of models to tackle problems currently addressed by involving experts and eliciting their judgments.

6. Modelling Measurements Errors with Auxiliary models

In this section, I argue that, although expert opinions should continue to be utilized to improve the measurement and prediction of international migration in Europe, they cannot replace the use of formal models. The question whether experts produce better predictions than models has occupied scientists for a long time. Armstrong (2001, 6.4), who for many years studied the use of expert judgments in forecasting cites "strong empirical evidence" that models (quantitative methods) are generally less biased and make more efficient use of data. To get more reliable and accurate expert information, Burgman (2016) advocates a change in attitudes towards expert estimates and predictions such that they are "treated with the same reverence as data, subjected to the same kind of cross-examination and verification." (Burgman 2016, vii). An expert's opinion is based on a model too: a mental model of true migration flows. Since the ultimate aim is to optimally combine quantitative methods (data and models) and qualitative methods (e.g., elicitation of opinions, expectations and predictions from experts, focus groups and stakeholders), formal models and mental models should be considered.

Mental models are outcomes of learning. Learning involves the development of mental models (cognitive schemes), which are representations of structured knowledge. Experts also use mental models and their beliefs and opinions are based on these models. Experts with more and better structured knowledge about a subject (better subject specialists) *and* with a strong empirical orientation are more likely to produce better estimates and predictions. When the expert's knowledge representation includes a deep insight in measurement procedures and the models that scientists use to produce estimates and predictions, the judgments may not be much different from the figures produced by the models that scientists use. An expert's degree of confidence in his or her estimates and predictions and his/her cognitive bias are influenced by the mental model. Initiatives to develop structured methods for elicitation, using well-defined protocols, are a first step to make explicit the mental models on which expert judgments are based.

Consider one of the measurement problems mentioned above, differences in duration threshold. Wiśniowski et al. (2013, 603) describe how they elicit from experts their opinions on the sensitivity of migration counts to duration thresholds, and how they translate that information into probability distributions to be used in estimations of migration flows. The expert opinions are translated into probability distributions via auxiliary variables $d^{(s)}$, which are assumed to follow log-normal distributions. It is not clear what substantive reasons exist for the selection of the log-normal. In this section, I show how correction factors can be obtained from a model of true migration flows. As a reference, I do not use the 12-month criterion, but a zero-month criterion (no time limit). I show that, for the same underlying data-generating process, that is, a process producing the true data, different results can be obtained depending on the measurement of the process. A measurement model that describes the impact of measurement method on the estimates of relocations was developed in a project to explore the use of micro-simulation for the harmonisation of migration statistics (Nowok 2010, Nowok and Willekens 2011).

Assume that people may relocate multiple times during an observation interval and that individuals act independently at the same constant relocation rate. This very simple situation is sufficient to illustrate the effects of differences in duration thresholds. Extensions will be considered at the end of this section. Relocations that satisfy these simple conditions are governed by a Poisson process. The distribution of numbers of relocations during an observation period is given by Equation (1). In Equation (1), λ is the expected number of relocations in a population during an observation period of length t. Since individuals relocate independently and at the same rate, we may consider the relocation of any single individual. The individual relocation rate is μ. It is the expected number of relocations experienced by an individual during a unit time interval, for example, one year. The expected number of relocations that the individual makes during a period of length t is μt. Define a migration as a relocation (change of usual residence) that is followed by a minimum duration of stay, the duration threshold. Migration statistics differ in the duration threshold used. Let d_m denote the duration threshold. An individual who relocates at time t is recorded as a migrant if he or she resides in the destination continuously for at least d_m years. Both actual and intended durations of stay may be used. The probability that an individual experiences n migrations between the onset of observation and time t if the duration threshold is d_m is:

$$\Pr\{N(t) = n|\mu, d_m\} = \frac{(\mu t z)^n}{n!} e^{[-\mu t z]} \tag{12}$$

where $z = e^{[-\mu d_m]}$ is the probability of no relocation within d_m years. It measures the proportion of relocations that are migrations, given the duration threshold d_m. The migration rate is $z\mu$.

The expected number of migrations during the interval of length t is

$$E[N_{d_m}(t)] = \mu t z = \mu t\, e^{[-\mu d_m]} \tag{13}$$

If a duration threshold of one year is used as a reference, as recommended by the United Nations, then a duration threshold of d_m results in a number of migrations experienced by an individual, that is O_{d_m} times the number of migrations experienced under the one-year

duration criterion, where O_{d_m} is (Nowok and Willekens 2011, 527):

$$O_{d_m} = E\left[\frac{N_{d_m}(t)}{N_{d_{12}}(t)}\right] = e^{[-\mu(d_m - d_{12})]} \tag{14}$$

The overestimation is $100\,(O_{d_m} - 1)$ %, with O_{d_m} measured in years. It is independent of the length of the observation period t. The overestimation is the same as the overcount factor, which is the percentage by which the number of migrations counted in a population is overestimated. For instance, if the relocation intensity is 0.2 and a country uses the six-month criterion, then $O_{d_m} = 1.10$, indicating that the reported figure overestimates the number of migrations by 10% measured in accordance with the UN guidelines (12-month criterion). Recording all relocations (no time limit) results in an overestimation of migrations by 22% (UN definition). Counting permanent migrations only, which are defined as migrations followed by a stay of at least five years (Nowok 2008), results in an underestimation of migrations (UN definition) by 55%. If 'permanent' means a stay of at least ten years, the underestimation is 84%, that is, only 16% of the migrations (UN definition) are recorded.

The overestimation is particularly sensitive to the relocation rate. The higher the rate, the higher the overestimation. Raymer et al. (2013, 807) reports that experts judge the number of migrations without time limit (i.e., relocation) to be about twice the number of permanent migrations (one-year criterion). If relocation is governed by a Poisson process with constant relocation rate, the relocation rate should be around 0.7 ($\mu = -ln(z) = -ln(0.5)$) to produce the expert judgment. That figure means that, on average, an individual relocates every 18 months, which is unrealistic. Another validity test of the Poisson model is to consider actual data on migration published by countries in Europe. Figures differ for a number of reasons listed above, with differences in the duration threshold being only one reason. Consider emigration from Poland. In the period 2002–2007, Poland registered an annual average of 22,306 emigrants to 18 EU and EFTA countries considered by De Beer et al. (2010), whereas the destination countries registered a total of 217,977 immigrants from Poland. Assuming that destination countries report immigrations correctly, the emigration rate in Poland would be six per thousand (in the period considered, the population of Poland was about 38 million). Poland records emigration if the person leaves the country permanently. Given the very low emigration rate of Polish residents, the Poisson model is unable to predict the large difference in recorded migrations in Poland and destination countries. The situation is worse if we consider the migration from Poland to one particular country. Consider migration to Sweden. During the same period, 2002–2007, Poland recorded an average annual emigration to Sweden of 303 persons, while Sweden recorded an annual average of 3,718 immigrants from Poland (De Beer et al. 2010). Sweden follows the UN guidelines in measuring migration. Given the very low rate of migration of Polish residents to Sweden, the Poisson model is unable to predict the large difference in recorded migrations from Poland to Sweden (Polish emigration data report only 8% of emigrants to Sweden if Sweden's immigration figures are considered accurate). The Poisson model could explain the difference if (a) the migration rate from Poland to Sweden is 0.2, (b) 'permanent' does not mean five or ten years, but a stay of at least 13.5 years, and (c) other measurement

errors have no effect. In that case, the measurement method used by Poland would underreport the true migration flow to Sweden by 92%.

The assumption that all individuals have the same relocation rate is not realistic. A large proportion of the population does not consider relocation and is therefore not really at risk of migration. Suppose that 2.5% of the residents of Poland consider emigration within a year. These have a much higher emigration rate than the average of the population of Poland (six per thousand). Since $0.006 = 0.975 \cdot 0 + 0.025m$, the emigration rate of people considering emigration is $m = \frac{0.006}{0.025} = 0.24$. Destination countries use different duration thresholds to measure immigration. If a duration threshold of six months is an acceptable average and 'permanent' emigration from Poland means a stay abroad longer than ten years, then the proportion of emigrants recorded in Poland is $e^{[-m(d_6-d_{12})]} = e^{[-0.24(10\ -\ 0.5)]} = 0.102$, which is 10%. That figure is a very good approximation of the proportion of emigrants recorded by Poland during the period 2002–2007. During the observation period, 1.7% of the emigrants from Poland emigrated to Sweden. Suppose residents of Poland have a slight preference for Sweden over other countries in the EU and EFTA region, increasing the emigration rate of those considering emigration to Sweden to 0.27 (instead of 0.24). That emigration rate results in a proportion of emigrants to Sweden recorded by Poland of $e^{[-0.27(10\ -\ 1)]} = 0.088$, which is the proportion observed in the period 2002–2007. The conclusion is that the Poisson model can yield an accurate estimate of the underreporting of emigration due to differences in duration threshold, if the migration rate does not apply to the total population but to the subpopulation that considers emigration, that is, the potential migrants. The relocation rate should apply to them and not to people who have no intention of emigrating or have an extremely low risk of emigration. To accurately describe underreporting or overreporting, the Poisson model should be extended to a mover-stayer model to incorporate unobserved heterogeneity in a population with respect to the desire to emigrate.

The experts, whose judgments were considered by Wiśniowski et al. (2013) and Raymer et al. (2013), indicate a much larger effect of the duration threshold than produced by the simple Poisson model. Table 1 shows the undercounts estimated by experts, the simple Poisson model with emigration rate of 0.24, and a mixture model, which is an extension of the mover-stayer model.

The expert judgments indicate that experts believe that onward or return migration soon after a previous migration is considerably higher than predicted by the Poisson model, which is a reasonable assumption. Unobserved heterogeneity may explain the deviation

Table 1. True migration flows (UN definition) as fractions of recorded flows. Expert judgments, Poisson model and mixture model.

Duration threshold	Experts judgment	Poisson model	Mixture model
No time limit	0.51	0.79	0.51
3 months	0.61	0.84	0.64
6 months	0.81	0.89	0.77
12 months	1.00	1.00	1.00
Permanent (p))	1.64		
5 years		2.61	1.80
10 years		8.67	2.98

from the Poisson model. Suppose a small proportion of the potential migrants ('movers') (for example, 6%) is very mobile and moves almost every six months (relocation rate is 1.8), while most (94%) potential migrants are modestly mobile and migrate every ten years, on average (migration rate is 0.1), then the true migration (according to the UN definition) as a fraction of the recorded flow is given in the third column of Table 1 (mixture model). The figures are close to the correction factors derived from the expert judgments. A model that distinguishes between people with and without a desire to emigrate produces true migration flows as fractions of the recorded flows that are similar to the expert judgments. If the population is indeed heterogeneous with respect to their desire to migrate, then part of the difference between the recorded migration flow and the true flow (according to the UN definition) can be attributed to the unobserved heterogeneity. That part is not a measurement problem caused by differences in duration thresholds, but a consequence of misspecification of the migration model. In that case, models of true migration flows, such as the one included in IMEM, should be extended to a mixture model to allow for that unobserved heterogeneity. A well-known example of a mixture model in the study of migration is the mover-stayer model. The prior probability distribution should also be a mixture distribution. Expert judgments on the proportion of movers in the population may be used to construct the mixture distribution.

An effect of population heterogeneity on relocation rates and migration rates is confirmed by the UNECE Task Force on analysis of international migration estimates using different length of stay definitions (UNECE 2012, 13f). The duration-of-stay dependence of relocation rates varies between males and females and between nationals and foreigners. The rate also varies between first international relocation and subsequent cross-border relocations. The first relocation is followed by a shorter duration of stay than subsequent relocations. The Task Force also presents, for different countries, the proportions of relocations using the three-month criterion that are recorded if the 12-month criterion is used. The findings differ greatly between countries.

7. Conclusion and Recommendations

In European countries, people feel uncomfortable with the level of immigration. Political parties that promise to regain control over immigration are on the rise. Politicians respond by discussing annual ceilings on the number of immigrants or a net number of migrants. How do they know the numbers? How valid and reliable are the numbers they use?

This article, as other articles on how we know the facts of international migration, paints a rather bleak picture of the state of international migration statistics. The problem was diagnosed almost 50 years ago and became acute when migration became the dominant component of demographic change and a major item on the political agenda. Several initiatives were taken at the national and European (and global) levels to improve the availability, quality and comparability of international migration statistics. The initiatives can be classified into broad categories. The first is the improvement of the production of migration statistics by the national statistical offices and other producers of official statistics in Member States, frequently in collaboration with Eurostat and in some cases with members of the European Statistical System in other Member States. It involves the documentation of the data collection process, the harmonisation of concepts and

measurement methods. In some cases, it also involves the use of mirror data supplied by other countries. The second category of initiatives is the development, by the research community, of statistical methods for estimating bilateral migration flows and for harmonizing available migration data. That often involves using different types of data from multiple sources.

That research produced a broad consensus among scientists that a dual strategy is required to improve statistics on international migration flow in Europe. The first component is that producers of statistics should thoroughly document the procedures they use to collect data and produce migration statistics. This documentation may be accompanied by a risk assessment, in which the types and sources of uncertainty in the data and the limitations in the production of statistics are made explicit. The second component of the dual strategy is oriented towards the research community. Models serve as a vehicle to effectively combine and integrate data from different sources and produce accurate and comparable migration estimates and migration statistics. The estimates are synthetic because data from different sources are combined and integrated. They yield harmonized statistics if the estimation procedure accounts for differences in the process of data collection and production of migration statistics. All steps of the estimation procedure should be thoroughly documented.

Past research on the estimation of international migration flows, reviewed in this article, revealed significant progress and a clear direction. A common element in all research is the use of migration flows by countries of immigration and flows by countries of emigration. The first such matrices were published in the mid 1970s, by the United Nations Economic Commission for Europe (UNECE). UNECE obtained the data from national statistical offices by a special request. Later, countries that collected the data provided the data annually. The data revealed that the immigration data and the emigration data are not consistent, that immigration is generally reported more accurately than emigration, and that some countries cannot produce such data or are not able to report immigration and emigration flow data on a regular basis. Initially, the focus of research was reconciliation of immigration and emigration flows. To make the data consistent, one set of country-specific adjustment factors was estimated and applied to the reported immigration data matrix and a different set was applied to the reported emigration data matrix (Poulain and Wattelar 1983). The correction factors are such that a measure of distance between the adjusted matrices is minimal, while some constraints imposed on the adjusted data are satisfied. The Euclidean distance was used initially, but later other distance functions were introduced. The adjusted immigration and emigration matrices are not equal. Poulain took the average of the two adjusted matrices. Abel (2010) gave priority to the correction factors for countries of immigration because immigration is generally measured more accurately than emigration. Initially, estimated migration flows were constrained to be equal to the total of the reported immigration flows. Later, additional constraints were imposed, but the basic approach remained constrained optimisation. De Beer et al. (2010) imposed the constraint that the adjusted immigration matrix and the adjusted emigration matrix have the same marginal totals.

Countries differ in the quality of their migration data and some countries do not report migration flows at all. Countries also use duration thresholds that may differ from the one-year duration of stay criterion recommended by the United Nations and Eurostat.

To account for the differences and to assure that the correction factors for countries with good data remain small, stepwise procedures were developed. The correction factors for countries with good data were estimated first, and those for countries with data limitations next. That introduced the need to judge the quality of the migration data reported by statistical offices. Expert judgments were solicited to rank countries by the perceived quality of their migration data. It also led to constraints on the correction factors for countries with good data and restrictions on the adjustments of some of the migration flows. Missing data constituted a separate problem. Some authors omitted countries with missing data. Abel (2010) estimated the missing flows using a spatial interaction model. He applied the EM (expectation-maximisation) algorithm to obtain the parameters of the model.

Raymer et al. (2013) adopted a similar idea, but replaced the constrained optimisation with a measurement model. A measurement model accounts explicitly for the sources of distortion in data due to differences in (1) concepts used, (2) measurements and data collection systems, in this case, differences in duration thresholds, (3) coverage of migrants, (4) undercount of migration, and (5) accuracy of the data collection mechanism. Raymer et al. also provided measures of uncertainty for all flow estimates and parameters in the model. A Bayesian approach is adopted to integrate the different types of data, covariate information, and prior knowledge. The migration model is used both to estimate the missing migration flow data and augment the measurement model. True migration flows that are consistent with the United Nations and Eurostat recommendation for the measurement of international migration (long-term migrations, i.e., migrations with duration threshold of 12 months) are treated as unobserved (latent) variables that need to be estimated from flow data reported by countries of immigration and countries of emigration, covariate information, and expert judgments. Wiśniowski et al. (2013) give a detailed account of how expert judgments are converted into prior distributions for subsequent use in the Bayesian inference.

The research community has followed an impressive trajectory in response to the call for migration data that are trustworthy and that can be used in migration governance and the migration debate. A milestone was EU Regulation No. 862/2007 of 11 July 2007 allowing Member States to use scientifically based and well documented statistical estimation methods in the production of migration statistics. The research community responded vigorously and produced the know-how and the technology to generate migration statistics that are harmonized and internally consistent, and accompanied by indicators of the accuracy of the statistics. That represents the state-of-the-art today. It is not the end of the trajectory. Further improvements are envisaged. The pace of improvements will critically depend on cooperation: cooperation among members of the European Statistical System and cooperation between the ESS and the research community. A concerted effort is needed to produce the evidence that allows a debate based on opinions *and* facts and motivates policies that are responsive to the evidence. The success of a concerted effort will depend on having a shared vision and a clear strategy. The vision is embedded in EU Regulation No. 862/2007: a combination of high-quality data collection and scientifically proven techniques provide the best guarantee for trustworthy international migration statistics. Since the sources of migration data that Member States rely on differ, the outcome will be a migration database that is synthetic, that is, which combines data from different sources. The database will evolve based on

stakeholders' changing expectations and queries, new sources of data and progress in science and technology. To master that process and find a proper balance between continuity and change, the perspective of a learning process is recommended. The synthetic database is a representation of reality. It represents a knowledge base for public debate, governance and research. New data may be incorporated ('assimilated') in the database without altering the structure of the database. When new data cannot be incorporated in an existing structure, the structure needs to be adjusted ('accommodation') which means that the model generating the database is updated. The information it contains should be reliable, but will not be perfect. Therefore, indicators of epistemic uncertainty (ignorance) and aleatory uncertainty (due to randomness) should be part of the database. Approaching the development of a synthetic database as a learning process paves the way for an effective use of insights from cognitive sciences and may guide the collection of new data.

The future trajectory involves several specific actions. Most have already been proposed and even advocated by others. The actions include:

1. Identify and document sources of data of international migration. The census and administrative records are main sources. Surveys, in particular household surveys, labour force surveys and designated migration surveys have untapped potential. Enhance migration mainstreaming in labour force surveys (e.g., migration questions and migration modules) and other data collection activities. Although the ESSC adopted a conceptual framework and work programme for migration statistics mainstreaming in 2010, it seems that guidelines and practical tools for mainstreaming migration in data collection have not yet been finalized. Gender mainstreaming may serve as a benchmark for migration mainstreaming (see e.g., European Commission 2017). Geolocation data generate new data sources for migration.
2. Statistical institutes that collect primary data or derive the statistics from primary data should publish detailed metadata on migration concepts and measures, and on the data collection process. The metadata should include a description of adjustment procedures introduced to account for nonreporting and other measurement problems. Scientists, who rely on metadata to develop methods for estimating and forecasting migration, should develop a thorough understanding of the migration data before engaging in estimation and/or forecasting (see also Disney et al. 2015).
3. Use mathematical/statistical models to produce the synthetic database. The distinction between migration model and measurement model (Raymer et al. 2013) is very useful. Migration models predict numbers of migrants by origin and destination, and by migrant attributes, such as sex, age, and education. Their policy relevance increases if they include the social and economic situation of migrants (Radermacher and Thorogood 2009). Measurement models should consider coverage, undercount, duration thresholds, accuracy, and other factors that cause observations to differ from true migration flows.
4. Include circular migration in models of migration. Duration thresholds considered in migration models should be flexible to cover permanent migration, short-term migration and circular migration. The modeling can benefit from the procedures developed by National Statistical Institutes for the measurement of short-term and

circular migration (see e.g., Johansson and Johansson 2016). UNECE and Eurostat support that development (see e.g., UNECE 2012, 2016; TEMPER 2015).

5. Develop life history models of migration, in addition to the population-level models in use today. Life history models adopt a longitudinal perspective and predict individual sequences of migrations/relocations and expected durations of stay in destination countries. They provide a logical way to incorporate lifetime migration (place of birth by place of residence), long-term migration, short-term migration, repeat migrations and circular migration in a single model in a coherent and consistent way. The UNECE Task Force on Measuring Circular Migration supports a life-history approach: "[*i*]n the ideal situation, the complete migration history of a person would be available. This would make it easy to determine whether a person qualifies as a circular migrant." (UNECE 2016, 20). Since data on individual migration histories will always be incomplete, truncation and censoring need to be dealt with (see also Beauchemin and Schoumaker 2016, 194). The theory of counting processes is the appropriate statistical theory for dealing with truncation and censoring (Aalen et al. 2008). Recently, DeWaard et al. (2017) used a life history model to estimate expected durations of stay in the EU-15 by migrants from new-accession countries.
6. Life history models may be extended to incorporate transnational activities. For instance, a person may obtain education in one country, work and raise children in another country and retire in a third. Activities are intertwined with migration. A temporary or circular migrant engages in more different activities than a permanent migrant. Life history models enable the integration of different types of relocations in the human life course and the assessment of how migration interacts with education, income generating activities, partnerships, the social network, and other aspects of life. Such an extension offers an analytical framework for the study of *multi-sited* individual and social lives (IOM 2010b).
7. Approach the development of the synthetic database as a *learning process*. A learning process builds representations of real world phenomena and improves the representations in light of new evidence and experiences. If one accepts that building a synthetic database is a learning process, then insights from cognitive science can help produce better data on migration.
8. The synthetic database is a step towards a smart or intelligent database. Databases may be trained to recognize data types, suggest estimation methods and signal new trends and discontinuities. The learning process may also point to individual decision processes and social processes that generate migration. Decision rules may be identified and incorporated in the database by replacing the statistical models by agent-based models (ABMs). Agent-based models simulate how agents process information (signals) from multiple sources in their environment and integrate that knowledge into a knowledge structure that is the basis for purposeful action (see Klabunde and Willekens 2016 for a review of agent-based models of migration and Willekens et al. 2016b and Klabunde et al. 2017 for recent illustrations).
9. Formalize learning. A formal method of learning that is particularly useful is the Bayesian method or Bayesian inference. A critical aspect of the Bayesian approach is to translate information or knowledge into probability distributions. Official

statisticians, who ultimately are responsible for developing the synthetic database, should be trained in the Bayesian method. Bayesian statisticians, on the other hand, should reach out to official statisticians and explain the logic of the Bayesian method.

10. Stimulate collaboration between National Statistical Institutes of sending and receiving countries to increase the comparability of migration data and enhance the harmonisation of data collection procedures and estimation methods. Promote exchange of data and the sharing of good practices. Secure adequate funding and training. UNECE (2010) developed guidelines on using data from destination countries to improve emigration statistics of origin countries (see also the UNECE site on migration statistics http://www.unece.org/stats/migration.html) and Eurostat established a secured web repository for exchanging migration data before their release (Kotowska and Villán Criado 2015). The 2016 New York Declaration for Refugees and Migrants also calls for enhanced international cooperation to improve migration data (United Nations General Assembly 2016).
11. Improve communication of migration data and publicize good practices. The Conference of European Statisticians' initiative to publish key recommendations and good practices in the communication of population projections shows the right direction (UNECE-CES Task Force on Population Projections 2016).
12. Bridge the gap between producers of statistics and scientists. Kotowska and Villán Criado, members of the European Statistical Advisory Committee, recommend that Eurostat takes the initiative and the lead to bridge that gap. Eurostat is, indeed, very well positioned and has demonstrated in the past decades that it can bring together scientists and producers of official statistics (Kotowska and Villán Criado 2015).
13. Methods for estimating emigration are particularly rare and should receive more attention. A very good point of departure is the report of the Suitland Working Group (Jensen 2013). Labour force surveys, household surveys and special migration surveys can be used to estimate rates of emigration. Wiśniowski (2017) uses Labour Force Surveys of Poland and the United Kingdom to estimate migration flows between the two countries. To identify emigrations, household surveys should collect data on the country of residence of household members living abroad, their age and the age at emigration. The sample design should assure enough observations to yield sufficiently precise estimates. Willekens et al. (2017) review the literature on the estimation of emigration. In addition, they use the Survey on Migration between Africa and Europe (MAFE) to estimate emigration rates from the Dakar region, Senegal to Europe, accounting for the complex sample design of the MAFE survey.
14. Produce reliable data on the number of irregular immigrants and integrate the data into the synthetic database. Reliable data on irregular migrants in the European Union do not exist. As border crossings by third country nationals are currently not registered, it is not possible to establish lists of overstayers. It is generally agreed that the majority of the 1.9 to 3.8 million of irregular immigrants within the EU overstay their Schengen visa (European Commission 2013), although this figure is not repeated in the 2016 version of the text (European Commission 2016b). The European Commission proposed the establishment of an advanced passenger information system for non-EU nationals travelling to the EU (European

Commission, 2013, 2016b). The system includes an Entry-Exit System (EES), with register of entries and exits, and a Travel Information and Authorization System (ETIAS). The system is modelled after the Electronic System for Travel Authorization (ESTA) and National Security Entry/Exit System (NSEERS) in the United States. The EES includes a mechanism to identify persons overstaying their authorized stay. In May 2015, a pilot EES project was started in Portugal. The system is believed to contribute to smart border management, but the experiences of the United States indicate the many challenges that emerge and need to be resolved.

15. Initiate and support a global, concerted effort to collect data on the root causes of international migration aimed at interventions that address emigration decisions and their motivating factors, rather than the consequences of the decisions. Several recommendations were made for a World Migration Survey building on the knowledge and experience gathered across the world in recent migration surveys of limited scale (see Section 3 of the article).
16. Expand research and analytical practice regarding measures of uncertainty for point estimates and related diagnostics for adequacy of the fit of the models employed.

8. References

Aalen, O.O. 1975. *Statistical Inference for a Family of Counting Processes*. PhD thesis. Berkeley: University of California.

Aalen, O.O., Ø. Borgan, and H.K. Gjessing. 2008. *Survival and Event History Analysis. A Process Point of View*. New York: Springer. Doi: http://dx.doi.org/10.1007/978-0-387-68560-1.

Abel, G.J. 2010. "Estimation of International Migration Flow Tables in Europe." *Journal of the Royal Statistical Society A* 173(4): 797–825. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00636.x.

Abel, G.J. 2013. "Estimating Global Migration Flow Tables Using Place of Birth Data." *Demographic Research* 28(18): 504–546. Doi: http://dx.doi.org/10.4054/DemRes.2013.28.18.

Abel, G.J. 2016. "Estimates of Global Bilateral Migration Flows By Gender Between 1960 and 2015". *International Migration Review* published online: August 13, 2018. Doi: https://doi.org/10.1111/imre.12327.

Abel, G.J. and N. Sander. 2014. "Quantifying Global International Migration Flows." *Science* 343(1520): 1520–1522. Doi: http://dx.doi.org/10.1126/science.1248676.

Andersen, P.K., Ø. Borgan, R. Gill, and N. Keiding. 1993. *Statistical Models Based on Counting Processes*. New York: Springer, ISBN: 978-0-387-94519-4. Doi: http://dx.doi.org/10.1007/978-1-4612-4348-9.

Armstrong, J.S. 2001. "Standards and Practices for Forecasting." In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, edited by J.S. Armstrong, Norwell, MA: Kluwer Academic Publishers (Springer): 679–732. ISBN: 978-0-7923-7930-0. Doi: http://dx.doi.org/10.1007/978-0-306-47630-3.

Azose, J.J. and A.E. Raftery. 2015. "Bayesian Probabilistic Projection of International Migration." *Demography* 52: 1627–1650. Doi: http://dx.doi.org/10.1007/s13524-015-0415-0.

Azose, J.J., H. Sevčíková, and A.E. Raftery. 2016. "Probabilistic Population Projections with Migration Uncertainty." *PNAS (Proceedings of the National Academy of Sciences of the United States of America)* 113(23): 6460–6465. Doi: http://dx.doi.org/10.1073/pnas.1606119113.

Beauchemin, C. 2013. *Statement Prepared for the Informal Hearings for High-level Dialogue on International Migration and Development (New York, July 15, 2013)*. Paris: International Union for the Scientific Study of Population (IUSSP). Available at: https://iussp.org/sites/default/files/IUSSP_Statement_UN_HLD_InformalHearings_InternationalMigration_10July.pdf (accessed February 2019).

Beauchemin, C. 2014. "A Manifesto for Quantitative Multi-sited Approaches to International Migration." *International Migration Review* 48(4): 921–938. Doi: http://dx.doi.org/10.1111/imre.12157.

Beauchemin, C. 2018. *Migration between Africa and Europe*. Cham: Springer. Doi: http://dx.doi.org/10.1007/978-3-319-69569-3.

Beauchemin, C. and B. Schoumaker. 2016. "Micro Methods: Longitudinal Surveys and Analysis." In *International Handbook of Migration and Population Distribution*, edited by M.J. White: 175–204. International Handbooks of Population 6. Dordrecht: Springer. Doi: http://dx.doi.org/10.1007/978-94-017-7282-2_9.

Betts, A. 2011. *Global Migration Governance*. Oxford: Oxford University Press. Doi: http://dx.doi.org/10.1093/acprof:oso/9780199600458.001.0001.

Bijak, J. 2011. *Forecasting International Migration in Europe. A Bayesian View*. New York: Springer.

Bijak, J. and J. Bryant. 2016. "Bayesian Demography 250 Years After Bayes." *Population Studies* 70(1): 1–19. Doi: http://dx.doi.org/10.1080/00324728.2015.1122826.

Bijak, J. and A. Wiśniowski. 2010. "Bayesian Forecasting of Immigration to Selected European Countries By Using Expert Knowledge." *Journal of the Royal Statistical Society A* 173(4): 775–796. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00635.x.

Bilgili, Ö. 2014. "Migrants' Multi-sited Social Lives." *Comparative Migration Studies* 2(3): 283–304. Doi: http://dx.doi.org/10.5117/CMS2014.3.BILG.

Bilsborrow, R.E., G. Hugo, A.S. Oberai, and H. Zlotnik. 1997. *International Migration Statistics: Guidelines for Improving Data Collection Systems*. Geneva: International Labour Organization. ISBN 9221095177.

Bilsborrow, R.E. 2016. "Concepts, Definitions and Data Collection Approaches." In *International Handbook of Migration and Population Distribution*, edited by M. White, 31–40. International Handbooks of Population 6. Dordrecht: Springer. Doi: http://dx.doi.org/10.1007/978-94-017-7282-2_7.

Bocquier, P. 2016. "Migration analysis using demographic surveys and surveillance systems." *International Handbook of Migration and Population Distribution*, edited by M. White. International Handbooks of Population 6. Dordrecht: Springer: 205–223. Doi: http://dx.doi.org/10.1007/978-94-017-7282-2_10.

Boswell, C. 2016. Report of the conference "Understanding and tackling the migration challenge: the role of research", Brussels, 4–5 February 2016. European Commission, CG Research and Innovation. Doi: http://dx.doi.org/10.2777/111442. Available at: https://ec.europa.eu/research/social-sciences/pdf/other_pubs/migration_conference_report_2016.pdf#view=fit&pagemode=none (accessed February 2019).

Brierley, M.J., J.J. Forster, J.W. McDonald, and P.W.F. Smith. 2008. "Bayesian Estimation of Migration Flows." In *International Migration in Europe. Data, Models and Estimates*, edited by J. Raymer and F. Willekens, 149–174. Chichester: Wiley. ISBN: 978-470-03233-6. Doi: http://dx.doi.org/10.1002/9780470985557.

Burgman, M.A. 2016. *Trusting Judgements. How to Get the Best Out of Experts*. Cambridge: Cambridge University Press. Doi: http://dx.doi.org/10.1017/CBO9781316282472.

Brooks, S., Gelman, A., Jone, G.I., and X-L. Meng (Eds.). 2011. *Handbook of Markov Chain Monte Carlo*. Boca Rotan: Chapman & Hall (Taylor & Francis). ISBN 9781420079418.

Caarls, K. 2016. NORFACE Research Programme on Migration. Migration in Europe: social, economic, cultural and policy dimensions. 2009–2014. Summary Report. Available at: https://pure.knaw.nl/portal/files/2869025/2016_Caarls_NWO_Summary_Report_NORFACE_final.pdf See also http://cordis.europa.eu/result/rcn/160917_en.html (accessed February 2019).

Cantisani, G. and M. Poulain. 2006. "Statistics on Population With Usual Residence." In *THESIM: Towards Harmonised European Statistics on International Migration*, edited by M. Poulain, N. Perrin, and A. Singleton: 181–201. Louvain-la-Neuve: Presses Universitaires de Louvain. ISBN 2-930344-95-4. Available at: http://www.seemig.eu/downloads/resources/THESIMFinalReport.pdf (accessed February 2019).

Cantisani, G., S. Farid, D. Pearce, and N. Perrin. 2009. *Guide on the Compilation of Statistics on International Migration in the Euro-Mediterranean Region*, Publication MEDSTAT II. Paris: Ed. ADETEF. Available at: http://www.unhcr.org/50a4f84d9.pdf (accessed February 2019).

Chater, N., J.B. Tenebaum, and A. Yuille. 2006. "Probabilisitc Models of Cognition: Conceptual Foundations." *Trends in Cognitive Sciences* 10(7): 287–291. Doi: http://dx.doi.org/10.1016/j.tics.2006.05.007.

Chebon, S., C. Faes, F. Cools, and H. Geys 2017. "Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex?" *Statistics in Medicine* 36(2): 345–361. Doi: http://dx.doi.org/10.1002/sim.7142.

Cohen, J., M. Roig, D.C. Reuman, and C. GoGwilt. 2009. "International Migration Beyond Gravity: A Statistical Model for Use in Population Projections." *PNAS (Proceedings of the National Academy of Sciences of the USA)* 105(40): 15269–15274. Doi: http://dx.doi.org/10.1073/pnas.0808185105.

Congdon, P. 1993. "Approaches to Modelling Overdispersion in the Analysis of Migration." *Environment and Planning A* 25(10): 1481–1510. Doi: http://dx.doi.org/10.1068/a251481.

Congdon, P. 2001. *Bayesian Statistical Modelling*. Chichester: Wiley. Doi: http://dx.doi.org/10.1068/a251481.

Courgeau, D. 1973. "Migrants et Migrations." *Population* 1: 95–129. English version published as Population Selected Papers No. 3, October 1979. Doi: http://dx.doi.org/10.2307/1530972.

Davies, R.B. and C.M. Guy. 1987. "The Statistical Modeling of Flow Data When the Poisson Assumption is Violated." *Geographical Analysis* 19(4): 300–314. Doi: http://dx.doi.org/10.1111/j.1538-4632.1987.tb00132.x.

De Beer, J., J. Raymer, R. van der Erf, and L. van Wissen. 2010. "Overcoming the Problems of Inconsistent International Migration Data: A New Method Applied to Flows in Europe." *European Journal of Population* 26: 459–481. Doi: http://dx.doi.org/10.1007/s10680-010-9220-z.

De Brauw, A. and C. Carletto. 2012. In *Improving the Measurement and Policy Relevance of Migration Information in Multi-Topic Household Surveys*. Washington D.C: Living Standard Measurement Study. World Bank. Available at: http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1199367264546/Migration_Data_v14.pdf (accessed February 2019).

DeWaard, J., K. Kim, and J. Raymer. 2012. "Migration Systems in Europe: Evidence from Harmonized Flow Data." *Demography* 49(4): 1307–1333. Doi: http://dx.doi.org/10.1007/s13524-012-0117-9.

DeWaard, J., J.T. Ha, J. Raymer, and A. Wiśniowski. 2017. "Migration From New-accession Countries and Duration Expectancy in the EU-15: 2002–2008." *European Journal of Population* 33(1): 33–53. Doi: http://dx.doi.org/10.1007/s10680-016-9383-3.

Disney, G. 2014. *Model-Based Estimates of UK Immigration*. PhD Thesis. Southampton: University of Southampton.

Disney, G., A. Wiśniowski, J.J. Forster, P.W.F. Smith, and J. Bijak. 2015. *Evaluation of Existing Migration Forecasting Methods and Models*, Report for the Migration Advisory Committee, ESRC Centre for Population Change. Southampton: University of Southampton. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/467405/Migration_Forecasting_report.pdf (accessed February 2019).

Dumont, J.-C. and G. Lemaitre. 2005. "Counting Immigrants and Expatriates in OECD Countries." *OECD Economic Studies* 3(1): 49–83. Doi: http://dx.doi.org/10.1787/eco_studies-v2005-art3-en.

European Commission. 2009. The Production Method of EU Statistics: A Vision for the Next Decade. Communication from the Commission to the European Parliament and the Council of 10 August 2009. COM(2009) 404 final. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:En:PDF (accessed February 2019).

European Commission. 2013. *Proposal for a Regulation of the European Parliament and of the Council Establishing an Entry/Exit System (EES) to Register Entry and Exit Data of Third Country Nationals Crossing the External Borders of the Member States of the European Union*. COM(2013) 95 final. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52013PC0095 (accessed February 2019).

European Commission. 2016a. *Research & Innovation Projects in Support of European Policy: Migration and Mobility*. Brussels: European Commission. DG Research and Innovation. EUR 27592 EN. Available at: http://ec.europa.eu/research/social-sciences/pdf/project_synopses/ki-na-27-592-en.pdf (accessed February 2019).

European Commission. 2016b. *Proposal for a Regulation of the European Parliament and of the Council Establishing an Entry/Exit System (EES) to Register Entry and Exit Data and Refusal of Entry Data of Third Country Nationals Crossing the External Borders of the Member States of the European Union and Determining the Conditions for Access to*

the EES for Law Enforcement Purposes and Amending Regulation (EC) No 767/2008 and Regulation (EU) No 1077/2011. COM(2016) 194 final. 2016/0106 (COD). 6th April 2016. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3 A52016PC0194 (accessed February 2019).

European Commission. 2017. *2017 Report on Equality Between Women and Men in the EU*. Brussels: European Commission, Directorate-General for Justice and Consumers. Doi: http://dx.doi.org/10.2838/52591.

European Statistical System. 2015. *ESS Vision 2020. Building the Future of European Statistics*. Luxembourg: Publication Office of the European Union. Available at: http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020 (accessed February 2019).

Eurostat. 2017. Households International Migration Surveys in the Mediterranean countries (MED-HIMS). Available at: http://ec.europa.eu/eurostat/web/european-neighbourhood-policy/enp-south/med-hims (accessed February 2019).

Evans, M., N. Hastings, and B. Peacock. 2000. *Statistical Distributions*. Third Edition. New York: Wiley ISBN: 0-471-37124-6 (Fourth edition: Forbes, C., M. Evans, N. Hastings and B. Peacock 2011. *Statistical distributions*. Hoboken, Jew Jersey: Wiley ISBN: 978-0-470-39063-4).

Fay, M.P. and S. Kim. 2017. "Confidence intervals for directly standardized rates using mid-p gamma intervals." *Biometrical Journal* 59(2): 377–387. Doi: http://dx.doi.org/10.1002/bimj.201600111.

Fiorio, L., G. Abel, J. Cai, E. Zagheni, I. Weber, and G. Vinué. 2017. "Using Twitter Data to Estimate the Relationship Between Short-term Mobility and Long-term Migration." In Proceedings of the 2017 ACM on Web Science (WebSci '17, Troy, New York). Doi: http://dx.doi.org/10.1145/3091478.3091496.

Flowerdew, R. and M. Aitkin. 1982. "A Method of Fitting the Gravity Model Based on the Poisson Distribution." *Journal of Regional Science* 22: 191–202. Doi: http://dx.doi.org/10.1111/j.1467-9787.1982.tb00744.x.

Frank, M. 2016. "Was Piaget a Bayesian? Analogies Between Piaget's Theory of Development and Formal Elements in the Bayesian Framework." Blog. Available at: http://babieslearninglanguage.blogspot.be/2016/04/was-piaget-bayesian.html (accessed February 2019).

Gerland, P. 2015. "Migration, Mobility and Big Data. An Overview." Paper presented at the GMG International Conference "Harnessing migration, remittances and diaspora contributions for financing sustainable development." New York, 26–27 May 2016. Available at: http://www.globalmigrationgroup.org/sites/default/files/GMG_2015_UNPD- PG_Migration.pptx (accessed February 2019).

Goodman, L.A. 1961. "Statistical methods for the mover-stayer model." *Journal of the American Statistical Association* 56(296): 841–868. Doi: https://doi.org/10.1080/01621459.1961.10482130.

Gopnik, A. and E. Bonawitz. 2015. "Bayesian Model of Child Development." *WIREs Cognitive Science* 6: 75–86.

Gopnik, A. and J.B. Tenenbaum. 2007. "Bayesian Networks, Bayesian Learning and Cognitive Development." *Developmental Science* 10(3): 281–287. Doi: http://dx.doi.org/10.1111/j.1467-7687.2007.00584.x.

Griffith, D.A. and R. Haining. 2006. "Beyond Mule Kicks: The Poisson Distribution in Geographical Analysis." *Geographical Analysis* 38: 123–139. Doi: https://doi.org/10.1111/j.0016-7363.2006.00679.x.

Hanea, A.M., M.F. McBride, M.A. Burgman, and others et al. 2016. "Investigate Discuss Estimate Aggregate for Structured Expert Judgement." *International Journal of Forecasting* 33(1): 267–279. Doi: http://dx.doi.org/10.1016/j.ijforecast.2016.02.008.

Hastings, W.K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57(1): 97–109. Doi: https://doi.org/10.1093/biomet/57.1.97.

Howson, C. and P. Urbach. 1989. *Scientific Reasoning. The Bayesian Approach*. La Salle, Illinois: Open Court Publishing Company. ISBN: 9780812695786.

Hughes, C., E. Zagheni, G. Abel, and others. 2016. *Inferring Migrations: Traditional Methods and New Approaches Based on Mobile Phone, Social Media, and Other Big Data. Feasibility Study on Inferring (Labour) Mobility and Migration in the European Union from Big Data and Social Media Data*, Report prepared for the European Commission, Programme for Employment and Social Innovation "EaSI" (2014–2020). Brussels: European Commission. Catalog N.:KE-02-16-632-EN-N. Available at: http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7910&type=2&furtherPubs=yes (accessed February 2019).

IOM. 2010a. "Migration and Transnationalism: Opportunities and Challenges. Background Paper for the Workshop 'Migration and Transnationalism: Opportunities and Challenges'." Geneva, 9–10 March 2010. Geneva: International Organization for Migration, International Dialogue on Migration. Available at: http://www.iom.int/idmtransnationalism (accessed February 2019).

IOM. 2010b. "Final Report of the Workshop 'Migration and Transnationalism: Opportunities and Challenges'." Geneva, 9–10 March 2010. Geneva: International Organization for Migration, International Dialogue on Migration. Available at: http://www.iom.int/idmtransnationalism (accessed February 2019).

IOM. 2011. *Glossary On Migration*. International Migration Law Series No. 25. Second edition, Geneva: International Organization for Migration (IOM). Available at: https://publications.iom.int/system/files/pdf/iml25_1.pdf (accessed February 2019).

Jacobs, R.A. and J.K. Kruschke. 2011. "Bayesian Learning Theory Applied to Human Cognition." *WIREs Cognitive Science* 2: 8–21. Doi: https://doi.org/10.1002/wcs.80.

Jensen, E.B. 2013. *A Review of Methods for Estimating Emigration*. Report of the Suitland Working Group on Migration Statistics. Washington D.C: Population Division, U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2013/demo/jensen-01.pdf (accessed February 2019).

Johansson, L. and T. Johansson. 2016. "Register for Mapping Circular Migration." Statistics Sweden Dokumenttyp 2015-12-14. Paper presented at the European Population Conference, Mainz, August 2016. Stockholm: Statistics Sweden. Available at: https://epc2016.princeton.edu/papers/160528 (accessed February 2019).

Kelly, J. 1987. "Improving the Comparability of International Migration Statistics: The Contribution of the Conference of European Statisticians from 1971 to Date." *International Migration Review* 21(4): 1017–1037. Doi: https://doi.org/10.1177%2F019791838702100406.

King, R. and A. Lulle. 2016. *Research on Migration: Facing Realities and Maximising Opportunities. A Policy Review*. Brussels: European Commission, DG Research and Innovation. Doi: http://dx.doi.org/10.2777/414370 (PDF). Available at: https://ec.europa.eu/research/social-sciences/pdf/policy_reviews/ki-04-15-841_en_n.pdf (accessed February 2019).

Klabunde, A. and F.J. Willekens. 2016. "Decision-making in Agent-based Models of Migration: State of the Art and Challenges." *European Journal of Population* 32(1): 73–97. Doi: https://doi.org/10.1007/s10680-015-9362-0.

Klabunde, A., S. Zinn, F. Willekens, and M. Leuchter. 2017. "Multistate Modelling Extended By Behavioural Rules: An Application to Migration." *Population Studies* 71(Supplement 1): S51–S67. Doi: https://doi.org/10.1080/00324728.2017.1350281.

Knauth, B. 2011. "Migration statistics mainstreaming." *Proceedings of the 58th World Statistical Congress 2011 (Dublin)*; The Hague: International Statistical Institute. Available at: http://2011.isiproceedings.org/papers/650162.pdf (accessed February 2019).

Kotowska, I.E. and I. Villán Criado. 2015. *Migration/migrants/integration*. European Statistical Advisory Committee, Doc 2015/1176 (manuscript).

Kraler, A. and D. Reichel. 2010. *Statistics on Migration, Integration and Discrimination in Europe. PROMINSTAT Final Report*. Available at: http://cordis.europa.eu/docs/publications/1243/124376691-6_en.pdf and http://www.prominstat.eu/drupal/node/64 (accessed February 2019).

Kraszewska, K., and D. Thorogood. 2010. *Migration Statistics Mainstreaming*. Working Paper, Joint UNECE/Eurostat Work Session on Migration Statistics, Geneva, 14–16 April 2010. Available at: http://www.unece.org/stats/documents/2010.04.migration.html#/ (accessed February 2019).

Kupiszewska, D. and B. Nowok. 2008. "Comparability of Statistics On International Migration Flows in the European Union." In *International Migration in Europe: Data, Models and Estimates*, edited by J. Raymer and F. Willekens, 41–72. Chichester: Wiley. ISBN: 978-470-03233-6. Doi: https://doi.org/10.1002/9780470985557.

Laczko, F. and M. Rango. 2014. "Can Big Data Help Us Achieve a 'Migration Data Revolution'?" *Migration Policy Practice (IOM)* 4(2): 20–29. ISSN 2223-5248. Available at: http://publications.iom.int/system/files/pdf/mpp16_24june2014.pdf (accessed February 2019).

Lanzieri, G. 2014a. "Filling the 'Migration Gaps' — Can Research Outcomes Help Us Improve Migration Statistics?" Prepared by the Statistical Office of the European Union (Eurostat) and presented at the Sixty-second plenary session of the Conference of European Statisticians, Paris, 9–11 April 2014. ECE/CES/2014/31. Available at: http://www.unece.org/stats/documents/2014.04.ces.html#/ and https://www.researchgate.net/publication/260096802_Filling_the_%27migration_gaps%27_-_can_research_outcomes_help_us_improve_migration_statistics (accessed February 2019).

Lanzieri, G. 2014b. "Test of an Estimation Method for Annual Migration Flows Between EU-EFTA Countries." Prepared by the Statistical Office of the European Union (Eurostat) and presented at the Sixty-second plenary session of the Conference of European Statisticians, Paris, 9–11 April 2014. Working Paper No. 9. Available at:

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2014/mtg1/WP_9_Eurostat.pdf (accessed February 2019).

Lanzieri, G. 2017a. "Summary Methodology of the 2015-Based Population Projections. Technical Note." Luxembourg: Eurostat ESTAT/F-2/GL. 3 March 2017. Available at: http://ec.europa.eu/eurostat/cache/metadata/Annexes/proj_esms_an1.pdf (accessed February 2019).

Lanzieri, G. 2017b. "Methodology for the Migration Assumptions in the 2015-based Population Projections. Technical note." Luxembourg: Eurostat ESTAT/F-2/GL. 5 July 2017. Available at: http://ec.europa.eu/eurostat/cache/metadata/Annexes/proj_esms_an3.pdf (accessed February 2019).

Liu, M.M., M.J. Creighton, F. Riosmena, and P. Baizán. 2016. "Prospects for the Comparative Study of International Migration Using Quasi-longitudinal Micro-data." *Demographic Research* 35(26): 745–782. Doi: https://doi.org/10.4054/DemRes.2016.35.26.

MEDSTAT Committee for the Coordination of Statistical Activities. 2011. *The HIMS Project (Household International Migration Survey)*. MEDSTAT Committee for the Coordination of Statistical Activities, 18th Session Luxembourg, September 2011. Available at: http://unstats.un.org/unsd/accsub/2011docs-18th/SA-2011-24-HIMS-Eurostat.pdf. See also the Eurostat website: http://ec.europa.eu/eurostat/web/european-neighbourhood-policy/enp-south/med-hims (accessed 9 February 2019).

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller 1953. "Equations os state calculations by fast computing machines." *Journal of Chemical Physics*, 21:1087–1092. Doi: https://doi.org/10.1063/1.1699114.

Miller, P.H. 1983. *Theories of Developmental Psychology*. San Francisco: W.H. Freeman and Co. ISBN: 0-7167-1432-9 (6th edition 2016 ISBN: 978-1429-278980).

Nowok, B. 2008. "Evolution of Migration Statistics in Selected Central European Countries." In *International Migration in Europe. Data, Models and Estimates*, edited by J. Raymer and F. Willekens, 73–87. Chichester: Wiley.

Nowok, B. 2010. *Harmonization By Simulation: A Contribution to Comparable International Migration Statistics in Europe*. Amsterdam: Rozenberg Publishers. ISBN: 978-90-367-4549-9. Available at: http://www.rug.nl/research/ursi/prc/research/popspace/migrationeurope?lang=en (accessed February 2019).

Nowok, B. and F. Willekens. 2011. "A Probabilistic Framework for Harmonisation of Migration Statistics." *Population, Space and Place* 17: 521–533. Doi: https://doi.org/10.1002/psp.624.

ONS. 2015. *Long-term International Migration Estimates. Methodology Document*. London: Office of National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/longterminternationalmigrationestimatesmethodology (accessed February 2019).

ONS. 2016. *Note on the Difference Between National Insurance Number Registrations and the Estimate of Long-term International Migration: 2016*. London: Office of National Statistics. 12 May 2016. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/noteonthedifferencebetweennationalinsurancenumberregistrationsandtheestimateoflongterminternationalmigration/2016 (accessed February 2019).

Payne, E.H., J.W. Hardin, L.E. Egede, V. Ramakrishnan, A. Selessie, and M. Gebregziabher. 2017. "Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling." *Statistical Methods in Medical Research* 26(4): 1802–1823. Doi: https://doi.org/10.1177/0962280215588569.

Perfors, A., J.B. Tenenbaum, T.L. Griffiths, and F. Xu. 2011. "A Tutorial Introduction to Bayesian Models of Cognitive Development." *Cognition* 120: 302–321. Doi: https://doi.org/10.1016/j.cognition.2010.11.015.

Pinsky, M.A. and S. Karlin. 2011. *An Introduction to Stochastic Modeling: Fourth edition*. Academic Press (Elsevier). ISBN 978-0233814162. Doi: http://doi.org/10.1016/C2009-1-61171-0.

Pisică, S. 2016. "Intra-EU Data Exchange – A Method for Improving Migration Statistics." Paper presented at the Eurostat conference "Towards more agile social statistics", Luxembourg, 28–20 Nov 2016. (Manuscript).

Poulain, M. 1991. "Un project d'harmonisation des statistiques de migration internationales au sein de la Communauté Européenne. (A project for the harmonisation of international migration statistics in the European Community)" *Revue Européenne des Migrations Internationales* 7(2): 115–138. Doi: https://doi.org/10.3406/remi.1992.1039.

Poulain, M. 1993. "Confrontation des Statistiques de Migrations Intra-Européennes: Vers plus d'Harmonisation? (Confronting the statistics on inter-European migration: Towards a greater harmonization?)." *European Journal of Population* 9(4): 353–381. Doi: https://doi.org/10.1007/BF01265643.

Poulain, M. 1999, Confrontation des statistiques de migration intra-européennes: vers une matrice complète?, Eurostat Working paper, n° 3/1999/E/N°5. Luxembourg: Eurostat. Available at: https://ec.europa.eu/eurostat/documents/3888793/5812665/KS-AP-01-016-DE.PDF/8e80d930-61c7-419a-9cec-34fa7343526a (accessed February 2019).

Poulain, M. 2008. "European Migration Statistics: Definitions, Data and Challenges." In *Mapping Linguistic Diversity in Multicultural Contexts*, edited by M. Barni and G. Extra, 43–66. Berlin/New York: Mouton de Gruyter. Doi: https://doi.org/10.1515/9783110207347.1.43.

Poulain, M. and L. Dal. 2008. *Estimation of Flows Within the Intra-EU Migration Matrix*. Report for the MIMOSA project. Available at: http://mimosa.cytise.be/Documents/Poulain_2008.pdf (accessed February 2019).

Poulain, M. and A. Herm. 2011. *Guide to Enhancing Migration Data in West and Central Africa*. Geneva: International Organization for Migration (IOM). Available at: http://publications.iom.int/system/files/pdf/dataguide_layout_101111.pdf (accessed February 2019).

Poulain, M. and A. Herm. 2013. "Central Population Registers as a Source of Demographic Statistics in Europe." *Population* 68(2): 215–247. Doi: https://doi.org/10.3917/popu.1302.0215.

Poulain, M. and C. Wattelar. 1983. "Les migrations intra-européennes: à la recherche d'un fil d'Ariane au travers des 21 pays du Conseil de l'Europe (Migration between the 21 countries members of the Council of Europe: Searching for the best estimation)."

Espace, Populations et Sociétés 1(2): 11–26. Doi: https://doi.org/10.3406/espos.1983.910.

Poulain, M., Perrin, N., and A. Singleton (Eds.). 2006. *THESIM: Towards Harmonised European Statistics on International Migration*. Louvain-la-Neuve: UCL Press. ISBN 2-930344-95-4. Available at: http://www.seemig.eu/downloads/resources/THESIM FinalReport.pdf (accessed February 2019).

Radermacher, W. and D. Thorogood. 2009. "Meeting the Growing Needs for Better Statistics on Migrants." Paper presented during DGINS Conference "Migration – Statistical Mainstreaming", Malta, 30 September – 1 October. Available at: http://ec.europa.eu/eurostat/documents/1001617/4339944/mainstreaming-w-radermacher.pdf/e3edaf 52-d5f6-471d-a26e-6f3e4a4516f0 (accessed February 2019).

Ravlik, M. 2014. *Determinants of International Migration: A Global Analysis*. Higher School of Economics Research Paper WP BRP 52/SOC/2014. Moscow: Higher School of Economic Research. Ravlik, Maria, Determinants of International Migration: A Global Analysis (October 2, 2014). Higher School of Economics Research Paper No. WP BRP 52/SOC/2014. Doi: http://dx.doi.org/10.2139/ssrn.2504441.

Raymer, J. (on behalf of the IMEM team). 2012. "Information Exchange and Modelling: Solutions to Imperfect Data on Population Movements." Economic Commission for Europe, Conference of European Statisticians, Work Session on Migration Statistics, 17–19 October 2012, Geneva. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2012/WP_14_USH_updated_by_R.pdf (accessed February 2019).

Raymer, J. and F. Willekens (Eds.). 2008. *International Migration in Europe. Data, Models and Estimates*. Chichester: Wiley. ISBN: 978-0-470-03233-6. Doi: https://doi.org/10.1002/9780470985557.

Raymer, J., A. Wiśniowski, J.J. Forster, P.W.F. Smith, and J. Bijak. 2013. "Integrated Modeling of European Migration." *Journal of the American Statistical Association* 108(503): 801–819. Doi: https://doi.org/10.1080/01621459.2013.789435.

Robert, C. and G. Casella. 2010. *Introducing Monte Carlo Methods with R*. New York: Springer. Doi: https://doi.org/10.1007/978-1-4419-1576-4.

Romanian Institute for Research on National Minorities. 2014. *National Policy Recommendations on the Enhancement of Migration Data for Romania*. Report Prepared in the Project 'SEEMIG Managing Migration and Its Effects – Transnational Actions Towards Evidence Based Strategies'. Available at: http://www.seemig.eu/downloads/outputs/SEEMIGPolicyRecommendationsRomania.pdf (accessed February 2019).

Schoorl, J.J., L. Heering, I. Esveldt, G. Groenewold, R. van der Erf, A. Bosch, H. de Valk, and B. de Bruijn. 2000. *Push and Pull Factors of International Migration. A Comparative Report*. Luxembourg: Eurostat. Available at: https://www.nidi.nl/shared/content/ output/2000/eurostat-2000-theme1-pushpull.pdf (accessed February 2019).

Schoumaker, B. and C. Beauchemin. 2015. "Reconstructing Trends in International Migration with Three Questions in Household Surveys: Lessons from the MAFE Project." *Demographic Research* 32(35): 983–1030. Doi: https://dx.doi.org/10.4054/DemRes.2015.32.35.

Skaliotis, M. and D. Thorogood. 2007. "Migration Statistics and Globalisation: Challenges for the European Statistical System." Paper presented at the 93rd DGINS Conference "The ESS response to globalisation. Are we doing enough?", Budapest, 20–21 September 2007. Available at: http://www.ksh.hu/pls/ksh/docs/eng/dgins/programme.html (accessed February 2019).

Statistics Netherlands. 2016. *Usual Residence Population Definition: Feasibility Study The Netherlands*. The Hague: Statistics Netherlands. Available at: https://www.cbs.nl/media/_pdf/2017/08/statistics-netherlands-feasibility-study.pdf (accessed February 2019).

TEMPER Team. 2015. *Report of the International Workshop on Methodological Challenges for the Study of Return and Circular Migration*. Event Review Series No. 1. Available at: http://www.temperproject.eu/wp-content/uploads/2015/06/Event-Review-1-2015.pdf (accessed February 2019).

Tourmen, C. 2016. "With or Beyond Piaget? A Dialogue Between New Probabilistic Models of Learning and the Theories of Piaget." *Human Development* 59: 4–25. Doi: https://doi.org/10.1159/000446670.

UNECE. 2010. "Guidelines for Exchanging Data to Improve Emigration Statistics." Paper prepared for the Task Force "Measuring emigration using data collected by the receiving country". Geneva: United Nations Economic Commission for Europe. Available at: http://www.unece.org/index.php?id=17456 (accessed February 2019).

UNECE. 2012. *Final Report of the UNECE Task Force on Analysis of International Migration Estimates Using Different Length of Stay Definitions*. UNECE, Conference of European Statisticians, 16th plenary session, Paris, 6–8 June 2012. Available at: http://www.unece.org/stats/migration/estimates.html.

UNECE. 2016. *Defining and Measuring Circular Migration. Prepared by the UNECE Task Force on Measuring Circular Migration*, ECE/CES/STAT/2016/5. Geneva: United Nations Economic Commission for Europe. Available at: http://www.unece.org/index.php?id=44717 (accessed February 2019).

UNECE-CES Task Force on Population Projections. 2016. *Key Recommendations and Good Practices in the Communication of Population Projections*. Geneva: Conference of European Statisticians, April 2016. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.11/2016/WP_28_TF_Report_v5_WithAppendix_.pdf (accessed February 2019).

United Nations. 1998. *Recommendations on Statistics of International Migration*. Statistical Papers Series M, No. 58, Rev. 1. New York: Statistics Division, United Nations Statistical Office. ISBN 92-1-161408-2. Available at: https://unstats.un.org/unsd/publication/seriesm/seriesm_58rev1e.pdf (accessed February 2019).

United Nations General Assembly. 2016. *New York Declaration for Refugees and Migrants*. UN General Assembly, A/Res/71/1, 3 October 2016. Available at: http://refugeesmigrants.un.org/declaration (accessed February 2019).

Valente, P. 2010. "Census Taking in Europe: How are Population Counted in 2010?" Population and Societies (Paris: Institut National d'Études Démographiques (INED)). No. 467. Available at: https://www.ined.fr/fichier/s_rubrique/19135/pesa467.en.pdf (accessed February 2019).

Van Dalen, H.P., G. Groenewold, and J.J. Schoorl. 2005. "Out of Africa: What Drives the Pressure to Emigrate?" *Journal of Population Economics* 18: 741–778. Doi: https://doi.org/10.1007/s00148-005-0003-5.

Van der Erf, R. and N. Van der Gaag. 2007. *An Iterative Procedure to Revise Available Data in the Double Entry Migration Matrix for 2002, 2003 and 2004*. Discussion Paper, Netherlands Interdisciplinary Demographic Institute, The Hague. Available at: http://mimosa.cytise.be/ (accessed February 2019).

Willekens, F. 1983. "Log-linear Modeling of Spatial Interaction." *Papers of the Regional Science Association* 52: 187–205. Doi: https://doi.org/10.1111/j.1435-5597.1983.tb01658.x.

Willekens, F. 1994. "Monitoring International Migration Flows in Europe. Towards a Statistical Data Base Combining Data from Different Sources." *European Journal of Population* 10(1): 1–42. Doi: https://doi.org/10.1007/BF01268210.

Willekens, F. 1999. "Modeling approaches to the indirect estimation of migration flows: From entropy to EM." *Mathematical Population Studies* 7(3): 239–278. Doi: https://doi.org/10.1080/08898489909525459.

Willekens, F. 2008. "Models of Migration: Observations and Judgements." In *International Migration in Europe. Data, Models and Estimates*, edited by J. Raymer and R. Willekens, 117–147. Chichester: Wiley. Doi: https://doi.org/10.1002/9780470985557.ch6.

Willekens, F. 2014. *Multistate Analysis of Life Histories with R*. Cham: Springer. ISBN: 978-3-319-08382-7. Doi: https://doi.org/10.1007/978-3-319-08383-4.

Willekens, F. 2016a. "Migration Flows: Measurement, Analysis and Modeling." In *International Handbook of Migration and Population Distribution*, edited by M. White. International Handbooks of Population 6, 225–241. Dordrecht: Springer. Doi: https://doi.org/10.1007/978-94-017-7282-2_11.

Willekens, F. 2016b. "The Decision to Emigrate: A Simulation Model Based on the Theory of Planned Behaviour." In *Agent-based Modelling in Population Studies: Concepts, Methods and Applications*, edited by A. Grow and J. van Bavel, 257–299. Dordrecht: Springer. Doi: https://doi.org/10.1007/978-3-319-32283-4_10.

Willekens, F. and J. Raymer. 2008. "Conclusion." In *International Migration in Europe. Data, Models and Estimates*, edited by J. Raymer and F. Willekens, 359–369. Chichester: Wiley. Doi: https://doi.org/10.1002/9780470985557.ch16.

Willekens, F., D. Massey, J. Raymer, and C. Beauchemin. 2016. "International Migration Under the Microscope." *Science* 352(6288): 897–899. Doi: https://doi.org/10.1126/science.aaf6545.

Willekens, F., A. Pór, and R. Raquillet. 1981. "Entropy, Multiproportional, and Quadratic Techniques for Inferring Detailed Migration Patterns from Aggregate Data." *IIASA Reports. A Journal of International Applied Systems Analysis* 4: 83–124. ISSN 0250-7625. Available at: http://pure.iiasa.ac.at/id/eprint/1561/1/IA-81-401.pdf (accessed February 2019).

Willekens, F., S. Zinn, and M. Leuchter. 2017. "Emigration Rates from Sample Surveys. An Application to Senegal." *Demography* 54(6): 2159–2179. Doi: https://doi.org/10.1007/s13524-017-0622-y.

Wiśniowski, A. 2017. "Combining Labour Force Survey Data to Estimate Migration Flows: The Case of Migration from Poland to the UK." *Journal of the Royal Statistical Society A* 180(Part 1): 185–202. Doi: http://dx.doi.org/10.1111/rssa.12189.

Wiśniowski, A., J. Bijak, S. Christiansen, J.J. Forster, N. Keilman, J. Raymer, and P.W.F. Smith. 2013. "Utilising Expert Opinion to Improve the Measurement of International Migration in Europe." *Journal of Official Statistics* 29(4): 583–607. Doi: https://doi.org/10.2478/jos-2013-0041.

Wiśniowski, A., J.J. Forster, P.W.F. Smith, J. Bijak, and J. Raymer. 2016. "Integrated Modelling of Age and Sex Patterns of European Migration." *Journal of the Royal Statistical Society A* 179(4): 1007–1024. Doi: http://dx.doi.org/10.1111/rssa.12177.

Yiu, S., V.T. Farewell, and B.D.M. Tom. 2017. "Exploring the Existence of a Stayer Population with Mover-stayer Counting Process Models: Application to Joint Damage in Psoriatic Arthritis." *Applied Statistics. Series C* 66(4): 669–690. Doi: https://doi.org/10.1111/rssc.12187.

Received March 2017
Revised February 2018
Accepted March 2018

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 279–283, http://dx.doi.org/10.2478/JOS-2019-0012

A Note on Dual System Population Size Estimator

Li-Chun Zhang[1]

Several countries are currently investigating the possibility of replacing the costly population census with a Population Data set derived from administrative sources, in combination with a purposely designed Population Coverage Survey. We formulate the assumptions of the dual system estimator in this context, and contrast them to the situation involving a census and a Census Coverage Survey.

Key words: Coverage error; undercount; capture-recapture; administrative data.

1. Introduction

The dual system estimator (DSE) has been used for adjusting population census undercoverage error with the help of a Census Coverage Survey (CCS). See Nirel and Glickman (2009) for a review. Wolter (1986) lists eight assumptions, that is, Assumption 1-6, 8, and 11, for the DSE in the most basic setting. Several countries are investigating the possibility of replacing the costly population census with a Population Data set (PD) derived from administrative sources, in combination with a purposely designed Population Coverage Survey (PCS). There is a need to pin down the assumptions of the DSE based on the PD and PCS, because the data generation process of a PD can be quite different and more difficult to model than that of a census. We propose to treat the PD as fixed and to consider the PCS as the only random source. This allows one to circumvent the problem of modelling the PD enumeration, where the mechanisms at the sources of the data may lie beyond the control of the statistician, and to focus on the design and implementation of the PCS, which is under the direct control of the statistician. In Section 2, we formulate the assumptions in the basic setting that is comparable to that of Wolter (1986). The advantages of the proposed conditional approach to the DSE, given the PD, will be explained in comparison to the traditional approach, where both the census and CCS are considered to be random. Some additional remarks on departures from the basic setting are given in Section 3, as well as some related ongoing developments.

2. The Four Assumptions for Consistency

Denote by $U = \{1, 2, \ldots, N\}$ the target population, which is of the unknown size N. Let A contain x enumeration records from the PD. Let S contain n records from the PCS. For each $i \in U$, let $\pi_i = E(\delta_i|A)$, where $\delta_i = 1$ if $i \in S$ and 0 otherwise. The notation $E(\cdot|A)$

[1] University of Southampton, Social Statistics and Demography, Highfield, Southampton, SO17 1BJ, United Kingdom. Email: L.Zhang@soton.ac.uk

emphasises that the enumeration in A or not is treated as fixed for all $i \in U$. In the most basic setting that is comparable to the DSE based on census and CCS, the following four assumptions are needed for the DSE based on PD and PCS:

(i) There are no duplicated records in A or S, and $A \cup S \subseteq U$.
(ii) The matched records between A and S can be identified without errors.
(iii) The capture probability in S is constant, that is, $\pi_i = \pi$ and $0 < \pi < 1$, for $i \in U$.
(iv) The captures in S are uncorrelated, that is, $Cov(\delta_i, \delta_j|A) = 0$ for $i \neq j \in U$.

According to (i), there are no duplicated or erroneous records in A or S, where a record i is erroneous if $i \notin U$. This is the same as Assumption 5 "Spurious Events" in Wolter (1986). The assumption (ii) combines Assumption 4 "Matching" and 7 "Nonresponse" in Wolter (1986). According to (iii), every population element has the same positive inclusion probability in the PCS. This is the same as Assumption 11 in Wolter (1986), except that it refers only to the capture probabilities in the PCS that are designed by the statistician, not the inclusion in the PD. Finally, the assumption (iv) is the same as Assumption 3 "Autonomous Independence" in Wolter (1986), except that it only pertains to the PCS enumeration, not the PD.

Let m be the number of matched records between A and S. Provided the assumptions (i) – (iii), we have $\mu_m = E(m|A) = \sum_{i \in A} \pi = x\pi$ and $\mu_n = E(n|A) = \sum_{i \in U} \pi = N\pi$, such that $x\mu_n/\mu_m = N$. Replacing μ_n and μ_m by n and m, respectively, we obtain the DSE

$$\hat{N} = xn/m. \tag{1}$$

It is important to notice that the only random source is the PCS that generates S, whereas we treat A (and x) as fixed, regardless of how complicated the data generation process may be that leads to the creation of A. In particular, the PD is allowed to have systematic undercoverage of the population, which is often the case with administrative registers. For instance, data set A may contain everyone that pays tax, but none who does not. The DSE (1) can still be motivated, because the estimated capture probability of the PCS among the tax payers, that is, m/x, can be extrapolated to the others, as long as the PCS satisfies assumption (iii). This provides additional flexibility, which is not permitted under the traditional approach to census and CCS. Moreover, treating A as fixed removes two of the three remaining assumptions of Wolter (1986). Assumption 2 "Multinomial" distribution of $(\delta_{i,census}, \delta_{i,CCS})$ is unnecessary, where $\delta_{i,census}$ and $\delta_{i,CCS}$ are the enumeration indicators of the census and CCS, now that inclusion or not in A is treated as fixed. Likewise, Assumption 8 "Causal Independence" between $\delta_{i,census}$ and $\delta_{i,CCS}$ is unnecessary, since the random variable δ_i cannot be correlated with a constant, that is, $i \in A$ or not.

Finally, Assumption 1 "Closure" of the population was used to ensure that the census and CCS aim at the same target population. In practice, it creates some tension to Assumption 8 "Causal Independence": to accommodate "Closure" one would conduct the CCS as close as possible to the census, yet doing so can potentially jeopardise "Causal Independence". The "Closure" assumption is no longer necessary, provided assumptions (i) and (iii) are satisfied. It is possible to implement any census population definition in the PCS, provided one can extract the data set A from the PD, which satisfies the assumption (i). For example, suppose the reference date is 11 November 2017 for a census night

population definition. The PCS may be deployed on the same day or immediately afterwards. Any member of the population has a chance to be enumerated in *S,* provided (iii), and no overcounting occurs, provided (i). Meanwhile, the processing of list *A* from the PD can take place both before and after 11 November 2017, aimed to satisfy assumption (i) and avoid erroneous enumeration. There is no need to assume that the target population itself is closed for a prolonged period after 11 November 2017.

Expanding $\hat{N}$ with respect to (n,m) around (μ_n,μ_m) yields

$$\hat{N}=N+\frac{x}{\mu_m}(n-\mu_n)-\frac{N}{\mu_m}(m-\mu_m)-\frac{x}{\mu_m^2}(n-\mu_n)(m-\mu_m)+\frac{N}{\mu_m^2}(m-\mu_m)^2+R_3, \quad (2)$$

where R_3 is the remainder. We have

$$\begin{aligned} E\left(\frac{\hat{N}}{N}|A\right)-1&=\left(1-\frac{x}{N}\right)\mu_m^{-2}V(m|A)+\frac{1}{N}E(R_3)\\ V(\hat{N})&\approx\frac{(N-x)^2}{\mu_m^2}V(m|A)+\frac{x^2}{\mu_m^2}V(n-m|A), \end{aligned} \quad (3)$$

where $V(m|A)=x\pi(1-\pi)$ and $V(n-m|A)=(N-x)\pi(1-\pi)$. Notice that we have used $Cov(n-m,m|A)=0$ and $Cov(n,m|A)=V(m|A)$, due to the assumption (iv). Provided $x/N=O(1)$ asymptotically, as $N\to\infty$, and $E(R_3)/N$ is of a lower order than the first term on the right-hand side of (3), we have $E\left(\frac{\hat{N}}{N}|A\right)\to 1$, because $V(m|A)/x=O(1)$ and $x/\mu_m=O(1)$. Now that $V\left(\frac{\hat{N}}{N}|A\right)\to 0$ in addition, the DSE (1) is such that $N/\hat{N}\to P_1$ asymptotically, as $N\to\infty$. The consistency of the DSE based on PD and PCS can thus be established under the assumptions (i) – (iv).

3. Additional Remarks

Below we consider potential departures from the four basic assumptions, taking them one by one in the reverse order.

Correlated PCS captures The assumption (iv) can be relaxed to allow correlated captures, such as intra-cluster correlated enumeration within the same household or building. Let the population U be partitioned into K clusters, denoted by $U=\cup_{k=1}^{K}U_k$.

(iv.c) $Cov(\delta_i,\delta_j)=0$ for $i\in U_k$ and $j\in U_l$, for $1\le k\ne l\le K$.

Provided (iv.c) instead of (iv), we have

$$E\left(\frac{\hat{N}}{N}|A\right)-1\approx\left(1-\frac{x}{N}\right)\mu_m^{-2}V(m|A)-\frac{x}{N}\mu_m^{-2}Cov(n-m,m|A),$$

where $V(m|A)=\sum_{i\in A}\pi_i(1-\pi_i)+\sum_{k=1}^{K}\sum_{i\ne j\in A_k}Cov(\delta_i,\delta_j)$, for $A_k=A\cap U_k$, and $Cov(n-m,m|A)=\sum_{k=1}^{K}\sum_{i\in A_k}\sum_{j\in A_k^c}Cov(\delta_i,\delta_j)$, for $A_k^c=U_k\backslash A_k$. Asymptotically, as $N\to\infty$, provided $x/N=O(1)$ as before, and $K/N=O(1)$ and $N_k=O(1)$, where N_k is the size of U_k which remains bounded asymptotically, the consistency of the DSE (1) is

retained. Moreover, the variance of $\hat{N}$ is now approximately given by

$$V(\hat{N}) = \frac{(N-x)^2}{\mu_m^2} V(m|A) + \frac{x^2}{\mu_m^2} V(n-m|A) - 2\frac{x(N-x)}{\mu_m^2} Cov(n-m, m|A),$$

where $V(n-m|A) = \sum_{i \in U \setminus A} \pi_i(1-\pi_i) + \sum_{k=1}^{K} \sum_{i \neq j \in A_k^c} Cov(\delta_i, \delta_j)$.

Heterogeneous PCS capture The assumption (iii) can be relaxed.

(iii.h) $\pi_i = \pi_h$ and $0 < \pi_h < 1$, for $i \in U_h$, where $U_1, \ldots, U_H$ form a post-stratification of the target population U.

Post-stratification is common in the practice of census-CCS DSE. Wolter (1986) introduces Assumption 7 "Post-stratification" to ensure that any variable used for the post-stratification is error-free. Provided this and the assumption (iii.h) instead of (iii), one may employ a post-stratified DSE based on the PD and PCS, which is given by

$$\hat{N}_p = \sum_{h=1}^{H} x_h n_h / m_h, \tag{4}$$

where x_h is the size of $A \cap U_h$, and n_h that of $S \cap U_h$, and m_h that of $A \cap S \cap U_h$. Asymptotically, as $N_h \to \infty$ for all $h = 1, \ldots, H$, we have $\hat{N}_p / N \xrightarrow{P} 1$.

(iii.a) $\bar{\pi}_A = \bar{\pi}_A^c$, where $\bar{\pi}_A = \sum_{i \in A} \pi_i / x$ and $\bar{\pi}_A^c = \sum_{i \in U \setminus A} \pi_i / (N-x)$ are the *average* capture probabilities among the population elements in and out of A, respectively.

According to (iii.a), the PCS does not have to achieve a constant capture probability across the population, which is less stringent than the assumption (iii). We have

$$x\frac{\mu_n}{\mu_m} = N\left(\frac{x}{N} + \left(1 - \frac{x}{N}\right)\frac{\bar{\pi}_A^c}{\bar{\pi}_A}\right) = N,$$

where $\mu_m = E(m|A) = x\bar{\pi}_A$, and $\mu_n = E(n|A) = N\left[(x/N)\bar{\pi}_A + (1-x/N)\bar{\pi}_A^c\right]$. The relative bias of the DSE is still given by (3), except that we now have $V(m|A) = \sum_{i \in A} \pi_i(1-\pi_i)$ and $V(n-m|A) = \sum_{i \in U \setminus A} \pi_i(1-\pi_i)$. Asymptotically, as $N \to \infty$, it converges to zero as before, so that the consistency property of the DSE (1) is retained.

(iii.ha) $\bar{\pi}_{A_h} = \bar{\pi}_{A_h}^c$, where $\bar{\pi}_{A_h} = \sum_{i \in A \cap U_h} \pi_i / x_h$ and $\bar{\pi}_{A_h}^c = \sum_{i \in U_h \setminus A} \pi_i / (N_h - x_h)$.

The assumption (iii.ha) combines (iii.h) and (iii.a), provided which the post-stratified DSE (4) retains its consistency property, as $N_h \to \infty$ for all $h = 1, \ldots, H$.

Linkage error The Matching assumption (ii) may be violated unless a unique identifier is available in both A and S, which can be used to link the records directly. See Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) for a discussion in the presence of linkage errors. To adjust the DSE, one needs to obtain estimates of the relevant linkage error probabilities, which is not an easy task in practice. Moreover, heterogeneous linkage error probabilities may further complicate the treatment of heterogeneous catch probabilities (in S). ONS-M8 (2013) outlines a potential alternative approach, which is to match A and S at

a cluster level (such as address or dwelling) that is not affected by linkage errors. However, the approach requires an additional assumption that the PCS fully enumerates everyone in the captured clusters, which may be difficult to satisfy in practice.

Erroneous enumeration The assumption (i) is violated for A if it contains erroneous records. The traditional approach is to include an additional survey (sampled from A) to estimate the over-coverage rate (e.g., Nirel and Glickman 2009). In some recent works, models and methods are developed to accommodate erroneous records directly. Zhang (2015) considers log-linear models of two PDs, subjected to both erroneous and missing records, together with the PCS. Zhang and Dunne (2017) apply the trimmed DSE to Irish data to explore the potential over-coverage error of the PD. In situations where the PD is compiled from multiple administrative registers, it is possible to trim one or more source registers directly. Di Cecco et al. (2018) develop latent class models based on four or more enumeration lists, all of which may be subjected to erroneous enumeration.

In particular, the treatment of linkage error and erroneous enumeration are important research topics for the census transformation programmes in the coming years.

4. References

Di Cecco, D., M. Di Zio, D. Filipponi, and I. Rocchetti. 2018. "Population Size Estimation Using Multiple Incomplete Lists with Overcoverage." *Journal of Official Statistics* 34: 557–572. Doi: http://dx.doi.org/10.2478/JOS-2018-0026.

Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158.

Di Consiglio, L. and T. Tuoto. 2015. "Coverage Evaluation on Probabilistically Linked Data." *Journal of Official Statistics* 31: 415–429. Doi: http://dx.doi.org/10.1515/JOS-2015-0025.

Nirel, R. and H. Glickman. 2009. "Sample Surveys and Censuses." In *Sample Surveys: Design, Methods and Applications, Vol 29A*, edited by D. Pfeffermann and C.R. Rao: 539–565.

ONS-M8. 2013. *Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory*. Avaiable at: https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/reportsandpublications.

Wolter, K. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://www.jstor.org/stable/2289222.

Zhang, L.-C. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://dx.doi.org/10.1515/JOS-2015-0023.

Zhang, L.-C. and J. Dunne. 2017. "Trimmed Dual System Estimation." In *Capture Recapture Methods for the Social and Medical Sciences*, edited by D. Böhning, J. Bunge, and P.v.d. Heijden: 239–259. Chapman and Hall/CRC.

Received November 2017
Revised April 2018
Accepted July 2018

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 285–286, http://dx.doi.org/10.2478/JOS-2019-0013

In Memory of Professor Susanne Rässler

Jörg Drechsler, Hans Kiesl, Florian Meinfelder, Trivellore E. Raghunathan, Donald B. Rubin, Nathaniel Schenker, and Elizabeth R. Zell

On the 29th of August 2018, Susanne Rässler, long-term associate editor of the Journal of Official Statistics, died far too early at the age of 55. All who knew her agree that she was a very special person. As Danny Pfeffermann said: "You could not just be a colleague. You had to be a friend."

Susanne received her PhD in Statistics from the Friedrich-Alexander Universität Erlangen-Nürnberg (FAU) in Germany in 1995 with a thesis comparing several estimators in unequal probability sampling. Following an offer to join the Department of Statistics and Econometrics at the FAU as a faculty member, she started her research in the field of statistical matching of multiple data sources, which culminated in her seminal book *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches* published in 2002. Two years later she temporarily left academia to accept a split appointment at the headquarters of the German Federal Employment Agency (BA) and the Institute for Employment Research (IAB), the research branch of the BA.

Despite working at the BA and the IAB for only three years, Susanne had a lasting impact at both institutions. At the IAB she established the Department for Statistical Methods and initiated numerous research projects on missing data, sampling designs, multiple imputation, and statistical disclosure control. Under her guidance, the IAB got involved in cutting edge research on methodological questions related to official statistics. As early as 2005, the institute experimented with split questionnaire designs to reduce the response burden of survey participants and to increase the quality of the gathered information. The institute was also the first statistical agency outside the United States to release multiply-imputed synthetic datasets to facilitate access to its highly sensitive establishment survey data for external researchers. Arguably, the most influential project initiated during her time at the IAB was a joint research project between the institute, the BA and Harvard University called TrEffeR (Treatment Effects and Prediction). The joint research enabled the BA for the first time to obtain detailed evaluations of its training measures based on the potential outcomes approach for causal inference (Rubin's Causal Model). The sophisticated procedures for matching treated and controls using the rich administrative databases available at the BA are still used today to evaluate all labor market programs offered by the Federal Employment Agency as well as for evaluating the relative effectiveness of the different providers of these programs.

In 2007 Susanne was appointed to the Chair of Statistics and Econometrics at the University of Bamberg, a position she held until her death. Because of her charming and warm personality, at Bamberg she was equally popular among students and academic colleagues, and her numerous committee positions within the faculty are testimony of her tireless commitment and dedication. In 2010 she initiated a new master's program in

Survey Statistics, which over the years evolved into one of the largest Statistics programs in Germany with currently over 100 enrolled students. In the same year, the University of Bamberg started the National Educational Panel Study (NEPS), a large scale multi-cohort longitudinal study with more than 60,000 participants. Susanne was the obvious choice for the head of the Statistical Methods Department of the NEPS and later became the scientific director of the same department, when the NEPS was integrated into the newly founded Leibniz Institute for Empirical Educational Trajectories (LIfBi).

Stimulating the exchange of ideas and fostering statistical literacy was always important for Susanne. She tirelessly worked to bridge communication gaps between academia, official statistics and the public. As part of this endeavour, she joined forces with the Bavarian State Office for Statistics and Data Processing and the IAB, and co-founded (as representative of the University of Bamberg) the Statistical Network of Bavaria in 2013. Even more important was her involvement in the last census in Germany. When the German statistical system started preparing for the 2011 Population Census (the first census in Germany since 1987), which was to feature novel methodology combining population registers and a large household survey, the German ministry of the interior established a scientific advisory body and invited Susanne to be a member of this board. In the years that followed, Susanne discussed methodological questions within the Board and with delegates of Destatis (the German national statistical institute). At the same time, she participated in numerous discussions on radio and on television, trying to convince the audience of the merits of the upcoming census. In 2017, some German states went to the Constitutional Court, claiming that parts of the census methodology were not in line with the constitution. Susanne served as a technical expert during the hearing, and may surely take the main credit for the court's 2018 decision in favour of the census.

Susanne was a wonderful mentor for her numerous PhD students. She always offered support, seeking any opportunity to promote the work of her students as well as to introduce them to the scientific community. The enthusiasm, which she showed for any proposed research idea (no matter how minor), motivated her students to keep working towards their degree even in times of little progress and much frustration.

For Susanne, all of her statistical collaborators were part of her family. She and her wonderfully supportive husband Hendrik would spend hours hosting the "statistical family" at their house. Many research ideas emerged from the discussions at the outdoor dining area, where Hendrik, though not a statistician, would always be there. Her warm and embracing nature and her remarkable energy will be deeply missed by all who had the fortune to know her.