Journal of Official Statistics
vol. 33, i. 4 (2017)

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 873–885, http://dx.doi.org/10.1515/JOS-2017-0040

The Morris Hansen Lecture

Hard-to-Survey Populations and the U.S. Census: Making Use of Social Marketing Campaigns

Nancy Bates[1]

Dann (2010, 151) defines social marketing as:

> *"The adaptation and adoption of commercial marketing activities, institutions and processes as a means to induce behavioral change in a targeted audience on a temporary or permanent basis to achieve a social goal".*

Social marketing campaigns have been used for decades in the United States as a means of influencing social behaviors. During World War II, the U.S. Government launched the Buy War Bonds campaign promoted by the War Advertising Council to encourage public participation in bond investment. More recently, health advocates have leveraged campaigns to influence behaviors including substance abuse prevention, family planning, HIV testing, and healthier food choices (Keller 2015; Andrews and Netemeyer 2015; Dholakia and Dholakia 2015; CDC 2017).

The U.S. constitution stipulates that a population census take place every ten years. In the 1960's, the census shifted from a personal-visit methodology to one that relied, in part, on self-response using census forms delivered by the United State Postal Service. The success of this shift depended upon the voluntary compliance of households to complete and return a questionnaire. In this context, social marketing can play a pivotal role in response rates by offering the U.S. Census Bureau targeted access to influence behaviors. In line with Dann's definition, social marketing can produce a temporary behavioral change of voluntary census participation (preferably by self-response); the targeted audiences are populations less inclined to participate (such as racial minorities and recent immigrants); and the social goal is an accurate census, which translates to better political representation, policy decisions, and community planning.

In the monograph *Hard to Survey Populations*, Tourangeau (2014) presents a framework of hard-to-survey populations according to the part of the survey lifecycle affected. For example, a group may be *hard to locate* because they are stigmatized or marginalized and prefer to remain hidden (e.g., undocumented immigrants or

[1] U.S. Census Bureau, Research and Methodology Directorate Room 5K140, 4600 Silver Hill Road, Washington DC 20233. U.S.A. Email: nancy.a.bates@census.gov

Acknowledgments: The author thanks Mary Mulry, Monica Vines, and Gina Walejko for comments on an earlier version of the article. Nancy is also grateful for the discussants of the 2016 Hansen lecture, Linda Jacobsen, and Brad Edwards. Disclaimer: The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

sexual minorities). Another group may be *hard to contact* because they are highly mobile, homeless, or have physical access barriers such as buzzer entries or reside in gated communities. Still other groups may be *hard to persuade* because they are suspicious of the federal government or have low levels of civic engagement and community attachment. Finally, other populations may be willing to participate in a census but *hard to interview* because of language barriers, low literacy, or lack of internet access.

Many of these barriers can be addressed by certain forms of social marketing. Populations that are hard to locate can be drawn out by engaging grassroots organizations – local community leaders and organizations recruited to serve as trusted voices to speak to constituents on behalf of the Census Bureau. Other elements of the campaign such as direct mail pieces, texts, and phone messages are a means to contact some hard-to-survey populations. Radio, print, and television paid advertisements with targeted messages (many in-language) are crafted and delivered to persuade some groups while click-to-complete digital ads make the actual interview process streamlined and tech-friendly.

The U.S. Census Bureau has a long history of leveraging social marketing techniques to promote the Decennial Census. The agency began a partnership with the Advertising Council (a pro-bono group of advertising agencies that create marketing campaigns for non-profit causes) beginning with the 1950 Census (U.S. Census Bureau, date unknown). Sports figures, politicians, and actors were featured in radio and television announcements informing populations about the benefits of census participation. By the 1990 Census, the agency realized the need to focus outreach activities more narrowly on racial and ethnic minorities as self-response rates between these groups and the general population began to diverge.

Up to and including the 1990 Census, community-level outreach activities and materials were developed by Census staff while national-level advertisements were delivered in the form of Public Service Announcements (PSAs). The creative content and production of the PSAs were donated to the agency by advertising agencies participating in the Ad Council. The placement of the spots in local and national markets was also free. However, because they generated no revenue, the census spots were run in suboptimal timeslots (after midnight through 5 am) and during programming with very low viewership. Consequently, the PSAs could not effectively target any particular hard-to-survey audience. Nonetheless, prior to the 1990 Census, these outreach efforts, along with earned media and high civic-mindedness of the United States population yielded self-response rates in line with agency expectations and budgets.

However, the 1990 Census proved to be a turning point. By late April 1990, the census was in crisis mode – the agency budgeted for a 70 percent self-response rate (five percentage points below the 1980 rate), yet the mail return was only 63 percent at the time personal-visit follow-ups were scheduled to begin (Bryant and Dunn 1995). The Census Director at the time, Dr. Barbara Bryant, had no choice but to request Congress allocate over 100 million additional dollars to complete the personal enumeration phase. This event motivated a change in social marketing campaigns for censuses to come – namely ones that involved contracts with advertising agencies to invest hundreds of millions of dollars to add and harness the benefits of *paid advertising* as part of future social marketing campaigns.

1. Operationalizing the Hard-to-Survey in the 2010 Census

The 2010 Census was the first to conduct a structured program of research to classify and pinpoint hard-to-survey populations for purposes of social marketing. One program of research produced a geographic audience segmentation of the entire United States population. This segmentation then became the backbone of the social marketing campaign, informing decisions from messaging, to partnership activities, to media spends, to the media channels selected to deliver the campaign messages.

Eight population segments were identified, each according to propensity for self-response. Three of the segments were predicted to have average or above average mail response – the other five had below average propensity and became the focal point for many aspects of the campaign. These hard-to-survey segments included two Ethnic Enclave groups, two Economically Disadvantaged groups, and a Single Unattached Mobile group (Bates and Mulry 2011). The Ethnic Enclave populations skewed toward recent immigrants with limited English proficiency living in urban areas often in larger households that included children. The Economically Disadvantaged were often made up of African American households with single mothers, lower education, and lower income. Finally, the Single Unattached Mobile segment was characterized by young, unmarried households prone to frequent moves.

In addition to the geographic segmentation, the agency sponsored a survey in 2008 to understand attitudes or "mindsets" of the population toward the Decennial Census. Of the five mindsets identified, three were classified as hard-to-survey and having a lower affinity towards the census. These included the Insulated, the Unacquainted, and the Cynical Fifth (Bates et al. 2009). The Insulated knew a little about the census but were generally indifferent toward it while the Unacquainted contained many foreign born who tended to be tenuously attached to their communities and completely unfamiliar with the census. The Cynical Fifth were identified as anti-government households who believed that census data could be used against them. Not surprisingly, the core characteristic of this population was resistance to responding to the census. The other two mindsets included the Leading Edge and Head Nodders. The former had high familiarity with the census and were predisposed to participate and advocate on behalf of the census. Similarly, the Head Nodders were positively predisposed, but also impressionable and vulnerable to negative news about the census. The behavioral mindsets and geographic segments were instrumental to the marketing campaign as means to develop and test messages that were believable and convincing with their intended targets. They also served as useful benchmarks to evaluate the effectiveness of the campaign as described below.

2. Components of the 2010 Census Social Marketing Campaign

The 2010 campaign was comprised of six interconnected components: paid advertising, earned media, local and national partnerships, the 2010 Census website, public relations, and the Census in Schools program. Paid advertising consisted of over 450 advertisements across television, radio, print, out-of-home and digital (Williams et al. 2015). Advertising was developed in twenty-eight different languages. Earned media and public relations included news releases, news conferences, blogs, a 2010 Census "Road Tour" (launched in early January on the network television Today Show), and a Take 10 Campaign that

broadcast daily mail participation rates by local areas. The partnership program worked to mobilize local leaders and advocacy groups to promote the census among their constituents. More than 257,000 governments, organizations, groups, and businesses partnered with the agency.

The 2010 campaign was likely the largest social marketing campaign in United States history. Between January and July 2010, the campaign was ranked among the top five advertisers with an average number of 42 campaign ad exposures with some targeted audiences having much higher "touches" (Williams et al. 2015). In fact, ad placements were so high that in some markets, the number of desired slots outstripped the available minority media inventory.

2.1. *Did Social Marketing Make a Difference?*

The 2010 campaign was extremely robust with countless interventions aimed to raise awareness, overcome barriers, and encourage participation. Extra partnership and advertising resources were distributed amongst the hardest-to-count areas, but only one controlled experiment was attempted to measure relationship between advertising "dosage" and behavior (see Bates et al. 2012). As such, attempts to quantify the effects of the campaign were extremely difficult. Nonetheless, I present several metrics that can be loosely construed as "proof" whether the campaign had a positive effect, particularly among hard-to-survey populations.

A common metric used to gauge the success of a census is the self-participation rate. In the case of the 2010 Census, self-response was achieved by completing and mailing back a paper questionnaire. Figure 1 depicts the mail self-response rates for the 1990, 2000, and 2010 Censuses. In 1990 (the last census to depend on a pro bono outreach campaign), the final mail response rate was projected at 70 percent but achieved only 65 percent. The 2000 Census (the first to use paid advertising), budgeted for a 61 percent response rate but achieved 67 percent. The 2010 Census (also with a paid campaign), also achieved a higher-than-projected mail response rate (projected was 64 percent with actual at 67 percent, see Fay et al. 1991; Letourneau 2012). While it is impossible to determine causation between the campaigns and levels of self-response, the fact that a longstanding

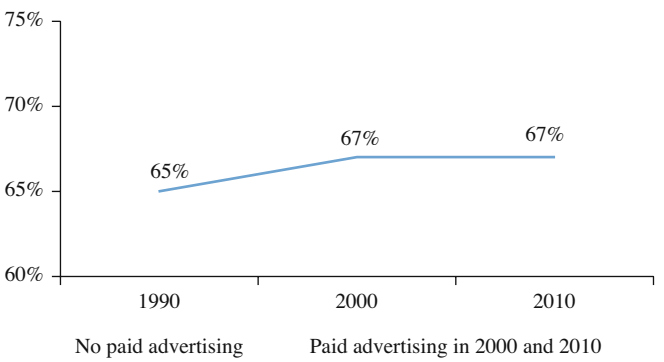

Fig. 1. US Census mail self-response rates: 1990–2010. Source: Fay et al. 1991; Letourneau 2012.

trend of declining self-response was reversed during the two censuses that engaged a paid campaign is noteworthy.

Another metric used to benchmark the effects of the Decennial marketing campaign are the mail check-in rates for the American Community Survey (ACS) during the time of the Decennial Census national campaign versus the same time period in a non-Decennial year. The ACS is a nationwide demographic survey sponsored by the U.S. Census Bureau that is continually in the field. The implementation method for the ACS in 2010 (mail prenotice, initial questionnaire, reminder postcard, and replacement questionnaire) was very similar to the method used in the 2010 Census. Bates and Mulry (2012) illustrate ACS mail check-in rates among the Economically Disadvantaged, Ethnic Enclave, and Single Unattached Mobile segments of the population for the March 2009 ACS panel (a non-decennial year) to the March 2010 ACS panel (the zenith of the 2010 Census marketing campaign). The ACS absolute mail check-in rates were higher for all segments in the decennial year compared to 2009 (Figure 2). However, the largest *percentage change* was documented among the five hardest-to-count segments. The two Ethnic Enclave segments had percentage changes of 17.4 and 29.3 between the campaign and no-campaign conditions; the two Economically Disadvantaged segments had percentage changes of 18.0 and 28.2; and the Single Unattached Mobile segment had a percentage change of 15.3 between the campaign and no-campaign conditions. Again, while the higher ACS self-participation cannot be directly attributed to the social marketing efforts, it is highly likely that the campaign played a significant role.

In 2010 much of the social marketing was directed toward the hardest-to-count segments including those skewing high on racial and ethnic minority households. Questions about census awareness and likelihood of self-responding were added to a daily Gallup tracking survey in January 2010 and continued until after Census Day, April 1, 2010. Figure 3 graphs, over time, the percent who reported having heard something about

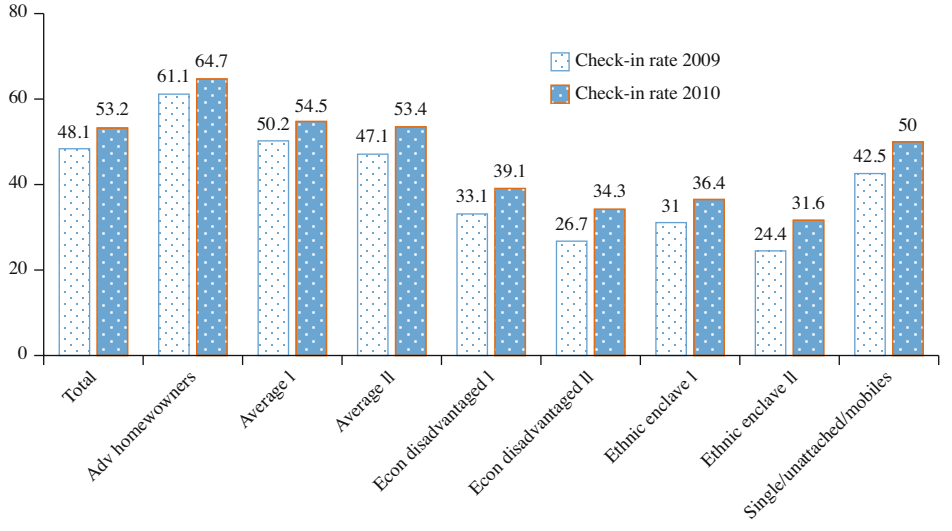

Fig. 2. ACS mail check-in rates by segment: non-Census year (2009) versus Decennial Census year (2010). Source: 2009 and 2010 ACS final mail check-in control file.

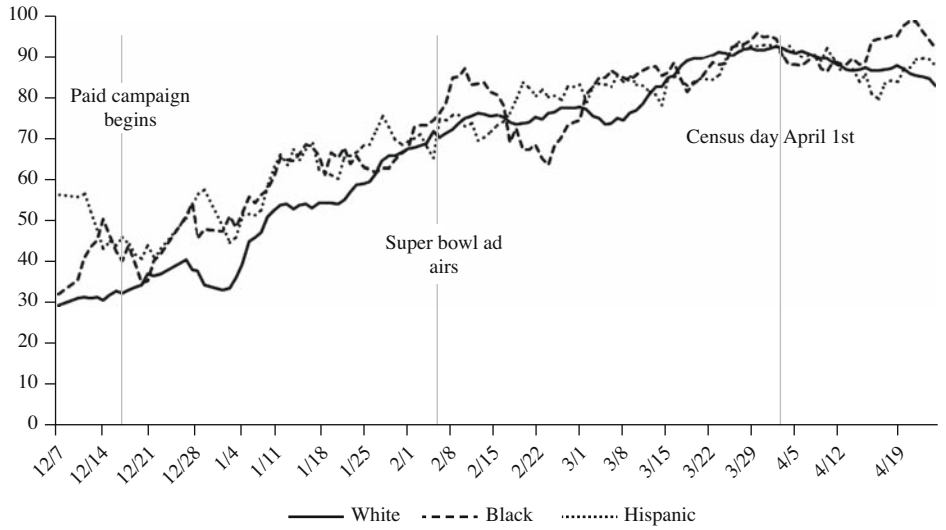

Fig. 3. "How much have you seen or heard recently-within the last week or so-about the 2010 Census?" Response: Heard great deal/some/a little Rolling Week %. Source: Miller and Wakejko 2010.

the 2010 Census. For much of the campaign the awareness among Blacks and Hispanics was higher than whites, suggesting success in raising awareness among some racial and ethnic minorities. Figure 4, however, suggests the campaign did not do as good a job raising awareness among the youngest 18–24 segment where the percent reporting having seen or heard something about the census consistently lagged below other age groups.

The last evaluation metric comes from data collected during a three-wave panel survey conducted as part of the larger independent campaign evaluation (U.S. Census Bureau 2012). In that panel, a subset of items used to form the original five mindsets were included. The first interview wave was conducted before the campaign began, the second wave occurred roughly mid-way through, and the third was fielded as the campaign was winding down. Using discriminant analysis, panel members were classified into mindsets at Wave 1, and then again at Waves 2 and 3 (Bates and Mulry 2012). This allowed us to track changes to mindsets over time as some respondents moved from one to another, due in part presumably, to exposure to the marketing campaign. Figure 5 illustrates the survey panel member's mindset distributions before, during, and at the end of the campaign. Of note is a decrease in the Unacquainted from eight percent at Wave 1 to less than one percent by Wave 3. Additionally, the Leading Edge (the mindset with the highest affinity toward the census), grew from 22 percent to 39 percent. Additionally, the Cynical Fifth decreased by roughly half from a pre-campaign 23 percent to 12 percent by the end. However, little change was observed among the size of the Insulated group.

3. Looking Forward: The 2020 Census Social Marketing Campaign

In August 2016, the U.S. Census Bureau announced a contract with the advertising agency Young and Rubicam (Y&R), to develop and deliver a social marketing campaign for the 2020 Census (U.S. Census 2016). In addition to Y&R, the contract includes the services of

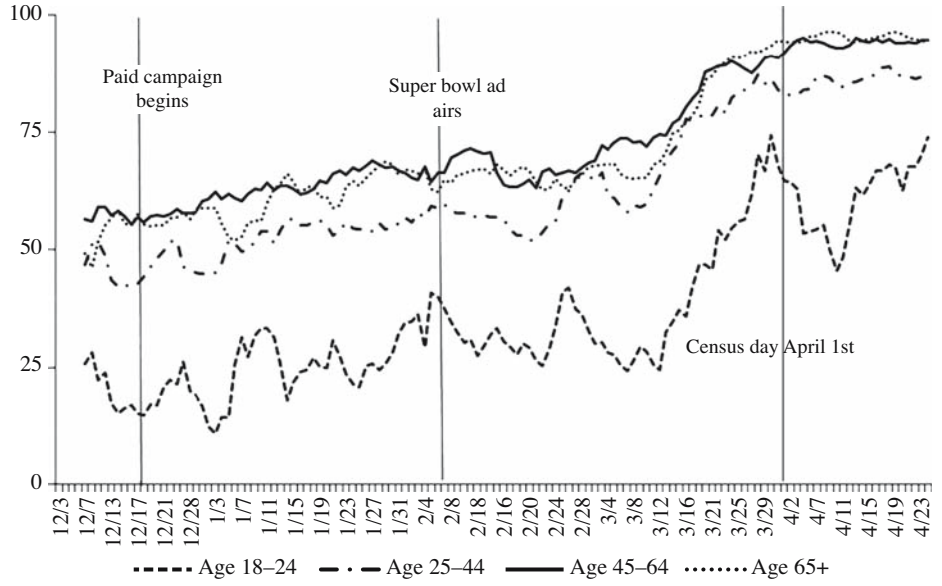

Fig. 4. "How likely are you to participate in the 2010 Census? By participate we mean fill out and mail in a Census form". Response: Definitely Will/Already Mailed Back Rolling Week %. Source: Miller and Wakejko 2010.

partner agencies each with expertise in reaching a specific population (e.g., African-Americans, veterans, Spanish speakers). Similar to the campaign in 2010, the 2020 campaign plans to make use of paid advertising. However, changes in the advertising industry coupled with the goal of maximizing self-response via the new Internet questionnaire necessitate changes from previous campaigns. For example, the campaign plans to deliver a much higher volume of paid advertising via electronic platforms including advertisements delivered to digital devices that allow households direct access to the census form simply by clicking an ad. These will include ads placed in search engines like Google and Bing, ads in social media feeds such as Facebook, and ads placed

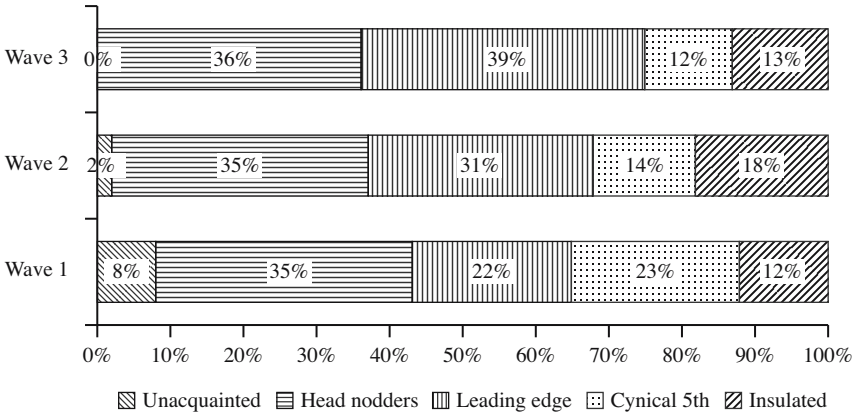

Fig. 5. Shift in mindsets over course of 2010 census social marketing campaign. Source: ICPCE panel survey data, 2010.

on targeted websites. This is in step with the growing percentage of advertising dollars allocated to digital media in the United States. In fact, in 2016 desktop and mobile advertising revenue surpassed television for the first time (Ad Age 2017).

Increased interest in digital advertising is correlated to the increase in mobile device ownership in the United States – in the context of hard-to-survey populations, this proliferation is important for several reasons. First, the smartphone ownership gap has closed between whites and racial and ethnic minorities in the United States. According to a Pew study (Rainie 2016), 66 percent of whites owned smartphones in 2016 compared to 68 percent of Blacks and 64 percent of Hispanics. Additionally, 34 percent reported using their phone as primary access to the internet, and this behavior was more likely among young adults, non-whites, and low income/low education populations. Digital ads that contain a direct link to the census form represent a new mode with potential to encourage online response among some populations less inclined to mail back a form or call a toll-free number.

A census test conducted in 2015 in the Savannah, Georgia Designated Market Area (DMA) provided the opportunity to experiment with the delivery of digital ads. In the United States, DMAs are geographic areas that share the same media markets. The Savannah DMA contains 17 counties in the Savannah Georgia area as well as three counties in South Carolina.

The test included a robust social marketing campaign including paid ads, hiring of local partnership specialists, and social and earned media. The digital ads featured embedded URLs enabling respondents to click on the ad and arrive at the landing page of the census test questionnaire. The test campaign delivered ads via search engines, social media in-feeds, and display ads on websites (U.S. Census 2017). In the Savannah test site, 90,000 households received mailings with an invitation to respond online. At the same time, advertisements with a URL were broadcast on television, radio, in print ads, on billboards, and in digital ads. Sampled mail households that had not responded within three weeks after the first mailing were mailed a paper questionnaire.

To understand the campaign's impact on hard-to-survey populations, we performed a Census-tract level exploratory Factor Analysis (FA) to identify hard-to-survey segments within the DMA (Virgile and Bates 2016). Inputs to the FA included predictors of Census 2010 self-response including variables such as age, poverty level, education, mobility, home ownership, race/ethnicity composition, and median household income. The FA identified two hard-to-survey segments. The first two contained tracts that skewed young adults who moved frequently and rented; the second tended to include African American female-headed households with low incomes and education. A third hard-to-survey segment was identified using data a different data source – the 2013 Federal Communication Commission (FCC) data file indicating number of households per 1,000 connected to residential high speed Internet. Tracts in the DMA containing between zero and 400 connected households per 1,000 were classified as "low internet connectivity". Because the Savannah test was primarily a test of online response, households in these tracts were classified as the third hard-to-survey segment. Combined, these three segments comprised approximately one-quarter (25.6 percent) of the Savannah DMA households.

Figure 6 illustrates the 2015 Savannah Census Test response mode distributions across the three segments, and overall. Internet response was the preferred choice among the young/mobile/renter segment (74.3 percent) but less so for the other two – 55.2 percent for

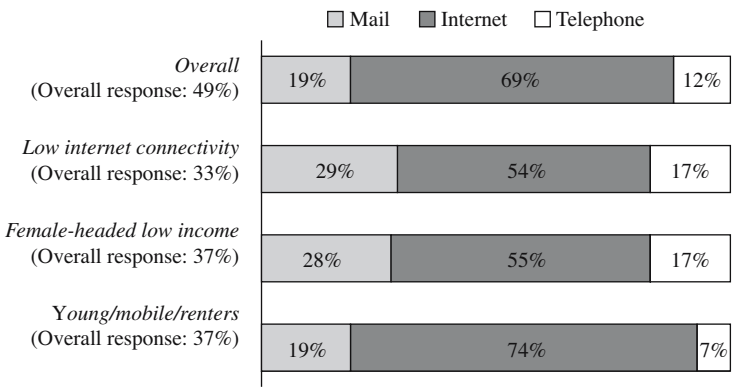

Fig. 6. 2015 Savannah Census Test: Response mode by hard-to-survey segments. Source: Virgile and Bates 2016.

the female headed/low income and education segment, and 53.7 percent for the low internet connectivity segment. Conversely, response by mail and telephone were above average for both the female-headed and low internet tracts (both at 17 percent).

Further analysis of online responses allowed insight into the campaign *source* responsible for driving households to the online form – for example, was it the direct mail piece, traditional ads such as television and radio, or digital ads? For all three hard-to-survey segments, the majority of online responses were most often generated from the URL advertised in the direct mail pieces (see Figure 7). However, close to one-quarter of online responses from the female-headed segment were the result of traditional ads (24.9 percent) and four to six percent of online responses from all three hard-to-survey segments were the result of clicking a digital ad. Given the relatively inexpensive cost of digital ads (USD2.39 was the highest spend per household in the Savannah test), even moderate rates of "click to respond" are noteworthy, particularly among populations that would otherwise require a costly personal visit follow-up in the Decennial Census.

In addition to embracing digital media and advertisements, the U.S. Census Bureau plans to leverage other innovations in the 2020 social marketing campaign. In 2020, households will have the option of responding online without the requirement of entering

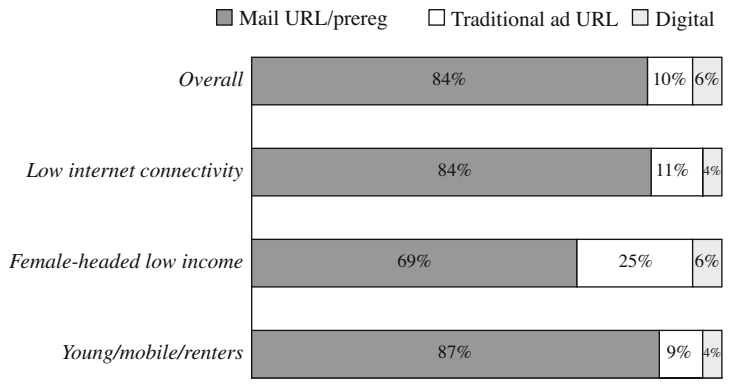

Fig. 7. 2015 Savannah Census Test: Source of online response among hard-to-survery segments. Source: Virgile and Bates 2016.

a pre-assigned address-based Census ID (included in the mail materials). This flexibility allows for online response by simply entering advertised URL's as part of the campaign or clicking on digital ads.

The U.S. Census Bureau will also conduct a survey to understand the current barriers, attitudes, and motivators to participate in the 2020 Census. These data – combined with ACS response data, previous census response data, and third party data such as voting behavior, internet connectivity, device ownership and computer usage – will be used to cluster households into audience segments with distinct attitudinal themes and media consumption habits. These "mindsets" will guide messaging and social marketing plans for different segments within the larger hard-to-survey population. These mindsets will help the creative advertising teams understand public perceptions of the 2020 Census and identify the type of communications most effective in motivating self-response.

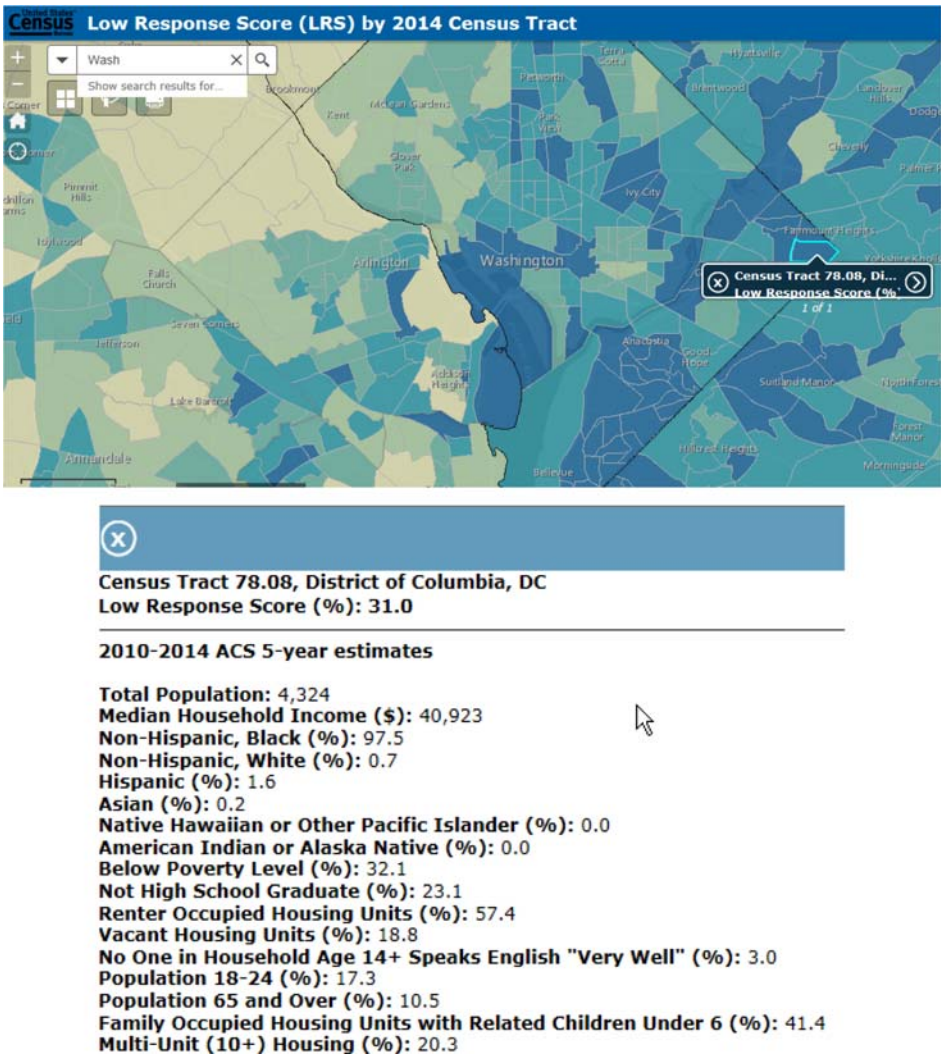

Fig. 8. Screen Shot of Response Outreach Area Mapper (ROAM) Web Application

The agency has also developed a new metric and mapping application to help locate hard-so-survey populations for Census 2020. The application displays Census tracts within user-defined geographic entities using a thematic color map to indicate each tract's Low Response Score or LRS. The LRS is the predicted *non self-participation* rate using Census 2010 mail response behavior as a guide (Erdman and Bates 2017). Users can enter various geographies including zip codes, states, places, and counties to identify tracts within the geography that have lower self-response propensities requiring extra attention and resources (darker colored tracts). In addition to displaying each tract's LRS, the application also displays a selected number of tract characteristics from the ACS five-year estimates (see Figure 8 for example screen shot of Washington DC and one tract in the Northeast quadrant). The application known as the Response Area Outreach Mapper (ROAM) is available for public use at (website: census.gov/roam) and expected to aid Census Partnership Specialists, city officials, elected officials, complete count committees, and community advocates alike.

The self-response target for Census 2020 is 63.5 percent, of which 47 percent is predicted to come from the Internet, five percent via telephone, and eleven percent through mail (U.S. Census Bureau 2017). Achieving these targets will undoubtedly hinge on the success of the social marketing campaign aimed at hard-to-survey populations. At the time of writing, public opinion of the federal government and U.S. Congress are at historical lows; undocumented populations are increasingly fearful of federal officials; and self-response to federal data collection continues to decline. Such challenges are top of mind to those planning the 2020 Census. Nonetheless, the agency is optimistic that some of these barriers can be overcome by a data-driven social marketing program.

4. References

Ad Age. 2017. "Desktop and Mobile Ad Revenue Surpasses TV for the First Time." April 26, 2017. Available at: http://adage.com/article/digital/digital-ad-revenue-surpasses-tv-desktop-iab/308808/ (accessed October 2017).

Andrews, J.C. and R.G. Netemeyer. 2015. "The Role of Social Marketing in Preventing and Reducing Substance Abuse." In *Persuasion and Social Marketing: Volume 3, Applications and Uses*, edited by D.W. Stewart, 155–194. California: Praeger.

Bates, N., F. Conrey, R. Zuwallack, D. Billia, L. Jacobsen, and T. White. 2009. "Messaging to America: Census Barriers, Attitudes and Motivators Survey (CBAMS)." In Proceedings of the 2009 American Statistical Association Survey Research Methods Section, Washington, DC, August 1–6. Available at: http://ww2.amstat.org/sections/srms/Proceedings/ (accessed October 2017).

Bates, N., K. McCue, and M. Lotti. 2012. "2010 Census Paid Advertising Heavy-Up Experiment Evaluation Report." 2010 Census Planning Memoranda Series No. 191, April 30. U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/2010census/pdf/2010_Census_PAHUE.pdf (accessed October 2017).

Bates, N. and M.H. Mulry. 2011. "Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse." *Journal of Official Statistics* 27: 601–618. http://www.jos.nu/Articles/abstract.asp?article=274601 (accessed October 2017).

Bates, N. and M.H. Mulry. 2012. "Did the 2010 Census Social Marketing Campaign Shift Public Mindsets?" In Proceedings of 2012 AAPOR conference in the American Statistical Association Survey Research Methods Section, 5257–5272, San Diego, CA, July 28–August 2. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2012f.html (accessed October 2017).

Bryant, B.E. and W. Dunn. 1995. *Moving Power and Money: The Politics of Census Taking*. New York: New Strategist Publications, Inc.

Centers for Disease Control. 2017. *#Doing It* Campaign: Testing for HIV. Available at: https://www.cdc.gov/actagainstaids/campaigns/doingit/index.html (accessed October 2017).

Dann, S. 2010. "Redefining Social Marketing with Contemporary Commercial Marketing Definitions." *Journal of Business Research* 63: 147–152. Doi: https://doi.org/10.1016/j.jbusres.2009.02.013.

Dholakia, R.R. and N. Dholakia. 2015. "Social Marketing and Family Planning: Family Planning Issues around the World." In *Persuasion and Social Marketing: Volume 3, Applications and Uses*, edited by D.W. Steward, 195–212. California: Praeger.

Erdman, C. and N. Bates. 2017. "The Low Response Score: A Metric to Locate, Predict, and Manage Hard-to-Survey Populations." *Public Opinion Quarterly* 81: 144–156. Doi: http://dx.doi.org/10.1093/poq/nfw040.

Fay, R., N. Bates, and J. Moore. 1991. "Lower Mail Response in the 1990 Census: A Preliminary Interpretation." In Proceedings of the Bureau of the Census 1991 Annual Research Conference, 3–32. Arlington, VA, March 1991. Available at: https://www.census.gov/srd/papers/pdf/rsm2010-13.pdf (accessed October 2017). 1991.

Keller, Punam A. 2015. "Social Marketing and Healthy Behavior." In *Persuasion and Social Marketing: Volume 3, Applications and Uses*, edited by D.W. Stewart, 9–38. California: Praeger.

Letourneau, E. 2012. "2010 Census Mail Response/Return Rates Assessment Report." 2010 Census Planning Memoranda Series No. 198. Available at: https://www.census.gov/2010census/pdf/2010_Census_Mail_Response_Return_Rates_Assessment.pdf (accessed October 2017).

Miller, P. and G. Walejko. 2010. *Tracking Study of Attitudes and Intention to Participate in the 2010 Census*. Technical Report to the U.S. Census Bureau, Contract YA1323-09-CQ-0032, Task Order 001.

Rainie, L. 2016. "Digital Divides 2016." Pew Research Center. Available at: http://www.pewinternet.org/2016/07/14/digital-divides-2016/ (accessed October 2017).

Tourangeau, R. 2014. "Defining Hard-to-Survey Populations." In *Hard to Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates, 3–20. Cambridge: Cambridge University Press.

U.S. Census Bureau. 2012. "2010 Census Integrated Communications Program (ICP) Evaluation Report". 2010 Census Planning Memoranda Series, March 15, 2012. Available at: https://www.census.gov/2010census/pdf/2010_Census_ICP_Evaluation.pdf (accessed October 2017).

U.S. Census Bureau 2016. Press release: "Census Bureau Statement on the Integrated Communications Contract for the 2020 Census". August 26th, 2016.

U.S. Census Bureau. 2017. 2020 Research and Testing: 2015 Census Test of Digital Advertising and Other Communications in the Savannah DMA Report. 2020 Census Program Internal Memorandum Series: <2017.14.i>.

U.S. Census Bureau. date unknown. Chapter 5: Census Promotional Program in *1990 Census of Population and Housing: History*. Washington DC. Available at: https://www.census.gov/history/pdf/1990proceduralhistory.pdf (accessed October 2017).

Virgile, M. and N. Bates. 2016. "Encouraging Online Response among Hard-to-Survey Populations: Digital Advertising and Influencer Calls." Paper presented at the annual conference of the American Association for Public Opinion Research, Austin, TX.

Williams, J.D., N. Bates, M.A. Lotti, and M.J. Wroblewski. 2015. "Marketing the 2010 Census: Meeting the Challenges of Persuasion in the Largest-Ever Social Marketing Campaign." In *Persuasion and Social Marketing: Volume 3, Applications and Uses*, edited by D.W. Stewart, 117–154. California: Praeger.

Received October 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 887–890, http://dx.doi.org/10.1515/JOS-2017-0041

Discussion

Brad Edwards[1]

Nancy Bates has pioneered an important subfield in survey methods: hard-to-survey populations. The topic of the Bates article in this issue is the U.S. Census' experience using social marketing to improve mail and web response rates for these populations. Populations can be hard to survey in many ways. The article is in line with a whole body of work Bates has developed on methods for asking about race, Hispanic origin, sexual orientation, and other characteristics, with a major focus on survey and census nonresponse. I was privileged to work with her on what became a major milestone in this field, an international conference in 2012 on hard-to-reach populations, and on the book that followed (Tourangeau et al. 2014). The conference was also the springboard for a special issue of the *Journal of Official Statistics* on the hard-to-reach (Willis et al. 2014).

The U.S. Decennial Census is a unique vehicle for studying the role social marketing can play in improving response from hard-to-survey groups because of its very large scale and its mandate to count every single person living in the United States. Other large scale vehicles, such as the U.S. presidential elections or national campaigns to change health behaviors, do not have such an exacting goal or such broad scope. Sample surveys are much smaller in scope and lack the resources to explore social marketing. The capability to target small areas like census tracts or even individuals in a sample at very low cost has only just emerged.

I work on surveys, not censuses, and for a contracting organization, not the U.S. Census. Most of my work has been on face-to-face surveys, not mail or mixed-mode like the Decennial Census. Nonetheless, there are many parallels in our work, and the program at the Census may become a model that brings social marketing into the forefront of survey methods.

The genesis of the research summarized in the Bates article comes from the observation that it is getting harder for censuses in the developed world to count a nation's people. There are many reasons for this, but changes in the concept of privacy is a prominent one. In the developed world, very few of us can live "off the grid". Countless electronic transactions mark our daily lives, and conventional notions of privacy have become outmoded. Almost everything about us seems public, but the notion persists that our personal data belong to us. It is "our" data. With that comes the recognition that our data is worth something to organizations, that we should get something for it, something like a discount at the supermarket, something greater than the satisfaction of contributing to the public good. It is not clear to many Americans anymore what they get from surveys and

[1] Westat, Inc., 1600 Research Boulevard Rockville, MD 20850-3195 Maryland 20009, U.S.A. Email: bradedwards@westat.com

censuses; neither is it clear what happens to it. When we give our data to organization A, ownership of the data is shared between us and A. We can then share our data with organization B, but A can also choose to share the data with another party, C, unless we have expressly forbidden it. It is easy to see how control – "ownership" – can be lost.

The late Eleanor Singer studied the relationship between privacy and willingness to participate in surveys. Her research showed that for most people, the decision to participate is the result of a cost-benefit tradeoff: What's in it for me? (Singer 2016, 2011, 2003). The length of time, the cognitive burden, the topic salience, the cash or non-cash incentive, what I've heard about it, what my friends and neighbors are likely to think about it, the interviewer's charm – all of those may be taken into account. But society's benefit, the common good, is less likely to enter into the equation than in the past. Trust in government, and in institutions like the press, have plummeted over the past decades.

Survey methodologists can manipulate some of these factors, of course (better publicity and advance letters, shorter interviews, larger incentives, interviewer training), but at a cost. Alongside response rate declines, we see costs increase. Many repeating surveys have spent increasing amounts of money to slow the response rate slide. (The Decennial Census from 1990 through 2010 is a prime example.) Because of small sample sizes in local areas, geographically targeted advertising has held little value for surveys: costs would be prohibitive, and advertising in a small geographic area could increase disclosure risk because it would be an indicator of the specific small second stage sampling units. What is so appealing about social marketing is that messages can be *tailored to individuals at very low cost*. Ideally the message prompts the individual to take action, and makes it very easy to act (by clicking a link in the message, for example, to reach the online census form).

The Census test of digital advertising in Savannah was a big leap down this road. It established that digital ads could drive hard-to-survey groups to online forms, increasing response rates and decreasing cost. This is the Holy Grail for survey methodologists. One can imagine many studies springing from it. Randomized control studies could test the effectiveness of different digital ads for different hard-to-survey groups; ad awareness and recall studies could explore the role that message exposure plays in the survey response decision; advertising metrics (reach, click-through rate, recall lift) could support comparative research on cost effectiveness of ad buys.

Underpinning the social advertising program at Census is the Low Response Score (LRS). It builds on the Hard to Survey (HTS) metric that Census developed for 2010. These response propensity models have stirred great interest for the survey world. Other organizations have been keen to apply them to gain efficiency in deploying resources, targeting ad campaigns and outreach materials, staging data collection efforts, and making weighting adjustments. The potential usefulness of the LRS in adaptive design is clear. It could help determine when phase capacity is likely to be reached, and when to change to a different approach. It could allow precise tailoring of each phase down to the block group or even the individual level for different population groups (personas, to use the Bates term).

Researchers at UCLA and Westat sought to use the HTS model as one of many data sources to explore the possibility of nonresponse bias on the California Health Interview Survey (CHIS), an annual telephone survey conducted for the State of California. Unexpectedly, no relationship was found between the achieved CHIS response rates and the HTS. (Lee et al. 2009) We speculated that, because the HTS was based on mail

response, it incorporated factors that were unique to mail, and did not include factors that may be unique to other modes. This speculation seemed to be supported by a more recent Westat analysis of the HTS and the LRS compared to achieved response rates on a large face-to-face data collection: again, no relationship. In 2017 Westat researchers completed a comparison of response rates on a *mail* survey with the HTS and the LRS, and again, found no relationship, concluding that those tools offer very little information to predict mail screener return on their study. Perhaps the mandatory nature of the Census invokes factors in propensity to respond that are not present in other mail surveys. We are currently building our own model for response to the face-to-face mode, based on our experience across a number of large household surveys. For face-to-face, gated communities and locked buildings are correlated with higher nonresponse rates. The impact of linguistically isolated communities is arguably also larger in face-to-face, because they are hard to identify and it is difficult to deploy interviewers on the ground who speak the language.

Bates' work is rooted in audience segmentation and social marketing. Running through this is the concept of social distance. The Census is trying to reach out to hard-to-survey audiences in ways that reduce the social distance, speaking to them in their language if you will. As the nearly two billion Facebook users know, suggestions from friends are very powerful, even if they are only "Facebook friends". The social media platform paradigm is built on "likes." It is an extension of an axiom from advertising: you will have more success if you begin with what the customer needs, not what you have to sell. This is the opposite of the survey paradigm, which start with the researcher's question, "What do I want to know?"

I was especially struck by the "cynical fifth" mindset in the Census segmentation. People with this mindset can have quite negative attitudes about the census or surveys or government, saying things like "It never does me any good," yet they may hold a basic belief that it *should* do something good, that it would be fair if only it could do something good, whatever that might be. This suggests the possibility of some common ground, where an ad campaign or an interviewer could turn the attitude around. For example, in an earlier draft Bates noted that many with this mindset have a strong belief in the U.S. Constitution; branding the census as something written into the Constitution might persuade them to respond.

A dark side travels with the rosy promise of social marketing for censuses and surveys, however. The success of social marketing depends on vast amounts of information being readily available, which strains the security apparatus to prevent unauthorized disclosure. One of the largest private data breaches ever detected was disclosed this summer. Personal identifying information for 143,000,000 adults was hacked from Equifax, one of three organizations that provide information about credit worthiness of individuals to lenders in the U.S. (White 2017). The potential for harm is great. The hacker could find a ready market on the dark web for these data, and its value could be retained for years into the future.

Planning for the 2020 decennial is in a critical stage. In the previous three decades, spending for the decennial has ramped up dramatically in the second year before the census, to levels two or three times the level of spending in the earlier years. At this writing, the government is operating on a continuing resolution, holding spending to the same level as last year, and Congress will be debating funding levels for Fiscal Year 2018

in the fall of 2017. With a requirement to spend less on the 2020 decennial than in 2010, the Census has embarked on an ambitious program of modernization, including a much more sophisticated social marketing campaign. The program requires an investment in new technologies and analytics. Without the funds now to ramp up and adequately test the program, the Census may have to fall back on more traditional methods, run out of funds, and risk missing millions of Americans who are hard to survey. A cynic might wish for that to happen, if those missed would be unlikely to vote in the cynic's favor. But that would torpedo one of the basic underpinnings of America's democracy, the principle of one person-one vote. Without an accurate count of persons, that principle is lost.

The U.S. Census has shown tremendous resilience over the past century, navigating some treacherous partisan waters and developing innovative methods that transform our field. It has never had a breach, and has never shared with anyone the decennial data it must guard for 70 years. It has a good brand. We can all hope that science and the public's common sense will prevail in these troubled times. But a stable budget and a supportive administration would help a lot. Better yet, move the Decennial Census out of the annual discretionary budget process, and onto a 10-year funding cycle.

References

Lee, S., E.R. Brown, D. Grant, T.R. Belin, and J.M. Brick. 2009. "Exploring Nonresponse Bias in a Health Survey Using Neighborhood Characteristics." *American Journal of Public Health* 99: 1811–1817. Available at: http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.2008.154161 (accessed 13 October 2017).

Singer, E. 2003. "Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits." *Journal of Official Statistics* 19: 273–286.

Singer, E. 2011. "Toward a Benefit-Cost Theory of Survey Participation." *Journal of Official Statistics* 27: 379–398.

Singer, E. 2016. "Reflections on Surveys' Past and Future." *Journal of Survey Statistics and Methodology* 4: 463–475. Doi: http://dx.doi.org/10.1093/jssam/smw026.

Tourangeau, R., B. Edwards, T.P. Johnson, K.M. Wolter, and N. Bates. 2014. *Hard-to-Survey Populations*. Cambridge University Press.

White, G.B. 2017. "A Cybersecurity Breach at Equifax Left Pretty Much Everyone's Financial Data Vulnerable." *The Atlantic*, September 7, 2017. Available at: https://www.theatlantic.com/business/archive/2017/09/equifax-cybersecurity-breach/539178/ (accessed October 13, 2017).

Willis, G.B., T.W. Smith, S. Shariff-Marco, et al. 2014. "Overview of the Special Issue on Surveying the Hard-to-Reach." *Journal of Official Statistics* 30: 171–176. Doi: http://dx.doi.org/10.2478/jos-2014-0011.

Received October 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 891–899, http://dx.doi.org/10.1515/JOS-2017-0042

Discussion

Linda A. Jacobsen[1]

In her Morris Hansen Lecture, Nancy Bates describes the Census Bureau's innovative approaches and success in identifying, reaching, and motivating hard-to-survey population groups with the 2010 Census social marketing campaign. She also previews the current plans for the 2020 Census campaign. In this discussion, I elaborate on several of the key challenges the Census Bureau – and the social marketing campaign – must address to successfully identify and reach hard-to-survey populations for the 2020 Census. These include: 1) The impact of changes in family formation processes and living arrangements; 2) Changes in the composition of the hard-to-survey population due to new census operations – principally the shift to Internet as the primary response option; and 3) The presence of both Internet and mail response options which will complicate messaging for the social marketing campaign. I also offer several options for the Census Bureau to consider in meeting these challenges.

1. The Impact of Changes in Family Formation Processes and Living Arrangements

For many years, adults in the United States followed a fairly uniform path in forming families. They first married, then began living with their spouse, and then had children shortly after marriage. They also tended to remain married throughout most of their adulthood – often remarrying after divorce or widowhood. But, significant increases in cohabitation and nonmarital childbearing over the past several decades have changed this process. Many adults first cohabit, then marry, and then have children, while others cohabit, have children, and then marry, and still others have children without cohabiting or marrying. Another important aspect of these changes has been the increase in relationship churn or repartnering – many adults today are serial cohabiters who have children with multiple partners without ever marrying. How dramatic have these changes been?

Today, a full 65 percent of women between the ages of 19 and 44 have ever cohabited. And, while cohabitation is higher among those with less education, almost three-fifths (58 percent) of women with a college degree have also cohabited (VanOrman and Scommegna 2016). Cohabiting unions are still fairly transitory, lasting about two to three years on average, and then transitioning to marriage or breaking up. In the early 1970s, only eleven percent of marriages were preceded by cohabitation, but by 2010 this share jumped to 69 percent (Manning and Stykes 2015). The trends in nonmarital childbearing are equally striking.

[1] Population Reference Bureau, U.S. Programs, 1875 Connecticut Avenue, NW, Suite 520, Washington DC 20009, U.S.A. Email: ljacobsen@prb.org

In the early 1980s, 21 percent of all births were nonmarital, and only six percent of births were to mothers who were cohabiting. By 2009–2013, more than 40 percent of all births were nonmarital, and one quarter were to cohabiting mothers**.** Nonmarital births are higher among racial and ethnic minority groups with Blacks having the highest share of nonmarital births at 75 percent, and Hispanics having the highest share to cohabiting mothers at 40 percent. And, while the exact estimates vary, researchers agree that the share of adults who have cohabited with more than one partner and who have children with multiple partners has been increasing, especially among women who have not completed college (VanOrman and Scommegna 2016; Monte 2017). These changes in family formation patterns have, in turn, caused important shifts in living arrangements.

The primary effect has been an increase in more complex household structures that include stepparents, stepsiblings, and half-siblings, as well as unrelated individuals. And, families often span multiple households. The increase in cohabitation and childbearing with multiple partners has affected the living arrangements of children in particular. Recent estimates indicate that more than 40 percent of children in the United States live in complex family households (VanOrman and Scommegna 2016). Yet, the current relationship question in both the Census and the American Community Survey (ACS) makes it difficult to identify and understand the relationships among the members of such complex households.

For the Census and ACS, respondents are instructed to list the name of a person living in the household who owns or rents the housing unit. This individual is called "Person 1", and the relationship question for all other household members identifies their relationship only to Person 1. The relationships between Person 1 and other household members in turn determine whether a household is considered to be a family or a nonfamily household. Although household structure has become more complex, this current relationship question prevents the Census Bureau (and researchers who use the data) from being able to determine if an adult other than Person 1 is the parent of a resident child, or how other household members are related to each other. This is particularly true in cohabiting couple households, where a child's classification as an "own" child or an "unrelated" child is arbitrary based on which unmarried partner is listed as Person 1. This problem is illustrated in Figure 1.

Figure 1 depicts three cohabiting couple households with a child. In Household 1, the child (C1) is the biological child of both unmarried partners (M1 and F1). Therefore, no matter which partner is listed as Person 1, the relationship question identifies this child as a "biological son or daughter" or as an "own child," and this household is classified as a family household. But, note that the relationship question doesn't allow us to determine whether F1 is the child's biological father. In Household 2, the child (C4) is the biological child of the female partner (M3), but not the male partner (F2). If the female partner is listed as Person 1 (as shown in Figure 1) then the child is again classified as an "own child" and this household is classified as a family household. For Household 2, we also can't determine whether F2 is the child's father. Household 3 in Figure 1 is identical to Household 2, except the male partner (F2) is listed as Person 1. Because the child (C4) is not the son or daughter of F2, the relationship question identifies C4 as an "unrelated child" and this household is classified as a nonfamily household. In this case, we can't determine whether M3 is the child's mother because the current relationship question does

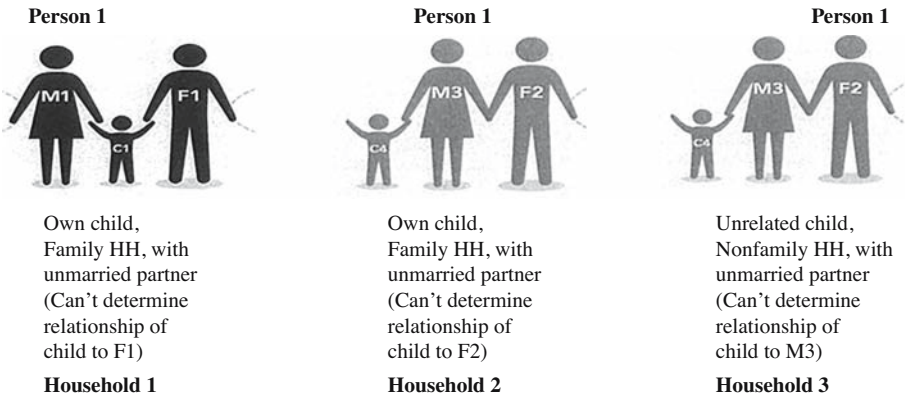

Fig. 1. Classification of Three Cohabiting Couple Households with a Child.

not identify the relationship(s) between a resident child and any other resident adult(s) other than Person 1. Although the second and third households depicted in Figure 1 have the same structure, they are classified completely differently depending on which unmarried partner is arbitrarily designated as Person 1.

Why do these changes in family formation and living arrangements matter for the 2020 Census? As living arrangements have become more fluid and transitory, individuals – particularly children – are more likely to split time between multiple households. As a result, it has become more confusing and challenging for respondents to understand decennial census and ACS residence rules and instructions defining who should be counted as a household member. In addition, these changes in living arrangements are more concentrated among groups who have historically been harder to survey – racial and ethnic minorities and those with less education who are economically disadvantaged. The confusion about who should be counted may not only impact response rates to the 2020 Census among complex households, but also differential undercount and the accuracy of the data.

Changing living arrangements are also important for 2020 because they are contributing to an increase in the net undercount of young children (ages 0 to 4) in the Census. This undercount for young children rose from less than two percent in 1980 to 4.6 percent in 2010 (O'Hare 2015), and Census Bureau research finds that unrelated children, children who are classified as "other relatives", and children living in complex households were more likely to be missed in the 2010 Census (U.S. Census Bureau 2014, 2017a, 2017b, and 2017c).

What options could the Census Bureau consider in response to this challenge? In the short-term, question(s) or pointers could be added to identify the relationship of children to resident adults other than Person 1. This is done in other surveys such as the Current Population Survey (CPS), and results in a more complete and accurate delineation of household composition and relationships. Although the primary response options for the Census and ACS (Internet and mail) are different from the CPS (in-person, telephone), it is important for the Census Bureau to develop and test changes to the relationship question for the decennial Census and the ACS to better address this rise in complex households and fluid living arrangements, especially among children.

Although the Census Bureau's extensive research on the undercount of young children has identified the types of households that may erroneously exclude young children as well as the characteristics of children who are more likely to be missed, it does not explain *why*. In the future, Census Bureau researchers might consider conducting studies that ask respondents **why** certain children were not included – especially those who are unrelated to or who are "other relatives" of Person 1. In the longer-term, the Census Bureau, and survey researchers in general, need to re-evaluate and evolve their concepts of residency as well as their instructions to respondents to better reflect the reality of current and future living arrangements.

2. Changes in the Hard-to-Survey Population for the 2020 Census

Another important challenge for the 2020 Census and social marketing campaign is accurately capturing potential shifts in the composition of the hard-to-survey population due to changes in census operations, particularly the switch to the Internet as the primary response option. Currently, the Census Bureau plans to use an Internet push option in the initial mailing for 80 percent of households, with only 20 percent initially receiving a paper Census form in the mail. This change in collection procedures may impact self-response rates for some groups. For example, those with historically high mail self-response rates – such as older adults – may be less likely to respond online. Similarly, those with historically low mail self-response rates – such as mobile young adults who are renters – may have higher Internet self-response rates. The Census Bureau implemented an Internet response option for the ACS in 2013, and some ACS research has shown that the switch to the Internet as the initial response option had a negative effect on self-response rates in some states, and for some population groups with lower Internet penetration, even when a paper form was mailed later (Baumgardner et al. 2014; Nichols et al. 2015).

While Bates cites data from a 2016 PEW study showing small differences in smart phone ownership between whites and racial and ethnic minorities, ACS data indicate that some historically hard-to-survey households are less likely to own computers or to have broadband Internet subscriptions at home. For example, in 2015 only 65 percent of non-Hispanic black households and 71 percent of Hispanic households had broadband (DSL, cable, fiber optic, mobile broadband, satellite, fixed wireless) subscriptions at home, compared with 79 percent of non-Hispanic white households (Ryan and Lewis 2017). Among households headed by someone age 65 or older, only 62 percent had a broadband subscription at home, and this drops to only 48 percent among householders who have not completed high school (Ryan and Lewis 2017). For 2020, it will be important for the Census Bureau to accurately identify, and for the social marketing campaign to reach, both new and historically hard-to-survey populations who are unable or unwilling to respond by Internet.

To facilitate the identification and geographic location of hard-to-survey populations, the Census Bureau developed a Low Response Score (LRS) based on Census 2010 mail response rates and data from the 2010 Census and the ACS (Erdman and Bates 2014). While Bates describes how the LRS for census tracts and a new mapping application will be used to help locate hard-to-survey populations for the 2020 Census, it is important to

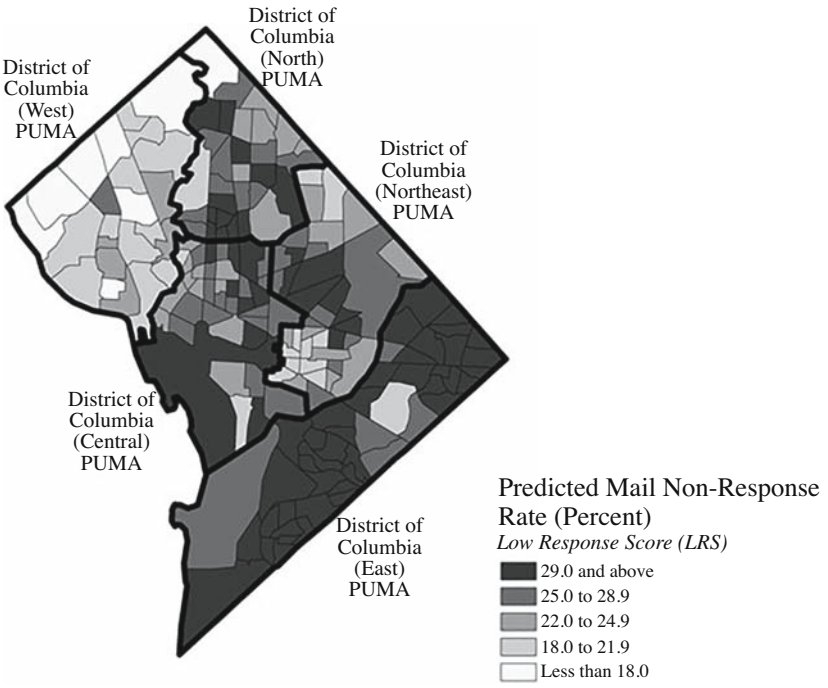

Fig. 2. Low Response Score by Census Tract for PUMAs in Washington, DC: 2010 Census and 2010–2014 American Community Survey.

recognize that the current LRS indicator may be less accurate for 2020 planning because it is based only on mail self-response rates. This potential problem can be illustrated by comparing tract-level LRS scores with recent data from the ACS on response mode, computer ownership, and Internet subscription.

Although the ACS is the largest annual sample survey in the United States, five years of ACS data must be combined to provide reliable estimates for census tracts and block groups. Since the Internet response option was not added to the ACS until 2013, ACS five-year estimates for 2013–2017 of Internet response, computer ownership, and Internet subscriptions will not be available until the fall of 2018. However, the ACS does provide one-year estimates for Public Use Microdata Areas (PUMAs) – the geographic areas included in the Census Bureau's Public Use Microdata Sample (PUMS) files. PUMAs are geographic areas within each state that contain at least 100,000 residents. Figure 2 shows the Census Bureau's 2016 LRS (predicted mail nonresponse rate) for all census tracts in Washington, DC, with the boundaries and labels for each of the five PUMAs overlaid. Tracts shaded with the darkest gray have the highest predicted mail nonresponse rate (29 percent or higher), while those shaded in lightest gray have the lowest predicted rate (less than 18 percent). Comparison of these tract-level LRS with data from the 2015 ACS one-year estimates highlights potential circumstances where the current LRS may be less accurate in predicting Internet response rates due to its reliance on Census 2010 mail response rates.

Table 1 provides response rates by mode for the District of Columbia and for each of its five PUMAs. Overall, about 42 percent of households in DC responded to the 2015 ACS

Table 1. Response mode, computer ownership, and Internet subscription for PUMAs in Washington, DC: 2015 American Community Survey.

	Percent of households responding by Internet	Percent of households responding by CATI or CAPI	Percent of households without a computer*	Percent of all households with a computer* but without an internet subscription
District of Columbia	**41.8**	**40.9**	**10.7**	**12.7**
West PUMA	57.4	24.5	2.7	3.6
North PUMA	35.7	45.1	13.7	11.2
Northeast PUMA	43.2	36.1	11.3	9.7
East PUMA	17.2	64.1	20.7	27.4
Central PUMA	54.4	33.1	5.6	9.3

Note: *Includes desktop, laptop, handheld (including smart phones), or other computers (including tablets). Excludes GPS devices and digital music players.

by Internet, but this varies from a low of only 17 percent of households in the East PUMA to a high of 57 percent in the West PUMA. The Internet response rates in the West and East PUMAs are consistent with their underlying tract-level predicted mail nonresponse rates. That is, most tracts in the West PUMA have low predicted mail nonresponse rates, while many tracts in the East PUMA have high rates. However, the Central PUMA has a large share of tracts with predicted high mail nonresponse rates, yet its 2015 ACS Internet response rate of 54 percent is almost as high as that in the West PUMA.

The percentage of households who responded to the 2015 ACS by Computer Assisted Telephone Interview (CATI) or Computer Assisted Personal Interview (CAPI) – more expensive modes of data collection than self-response by mail or internet – is generally consistent with the tract-level LRS, particularly in the East and West PUMAs. Almost two-thirds (64 percent) of households in the East PUMA responded by CATI or CAPI, compared with only one-fourth (24.5 percent) of households in the West PUMA (see Table 1).

Analysis of 2015 ACS data on household computer ownership and presence of an Internet subscription help to explain the response mode patterns by PUMA. Overall, about eleven percent of households in DC do not have a desktop, laptop, handheld computer (includes smart phones), tablet, or other computer, compared with about 13 percent of households nationwide. This share is very low among households in the West PUMA (3 percent), but jumps to more than one-fifth of households (21 percent) in the East PUMA (see Table 1). The percentage of households without a computer is also very low in the Central PUMA (six percent), even though many of its underlying tracts have high predicted mail nonresponse rates.

Of course, it is not only computer ownership that is important for Internet response, but also access to the Internet. Table 1 shows that about 13 percent of all households in DC have a desktop, laptop, handheld computer (includes smart phones), tablet, or other computer, but do not have a subscription to the Internet (payment for a type of service that provides Internet access such as a data plan for a mobile phone, a cable modem, DSL, or other). These shares are again lowest among households in the West and Central PUMAs,

but more than one-fourth (27 percent) of households in the East PUMA have some type of computer but no Internet subscription at home.

The demographic characteristics of households in the Central PUMA help to explain the discrepancy in the tract-level LRS and high Internet response rates in the 2015 ACS. Central PUMA residents match the profile of young, single, mobile renters who are more likely to respond online than by mail. Although this is just one example, it indicates that the current LRS may provide less accurate predictions of Census 2020 self-response rates for geographic areas with population groups who are either more likely or less likely to respond by Internet than by mail. The Census Bureau's plan to adjust the current LRS to include multiple response modes (e.g., Internet and mail) is an important enhancement that will increase the accuracy and utility of the LRS in locating hard-to-survey populations for Census 2020.

3. Multiple Response Modes Will Complicate Messaging for 2020

For the 2020 Census, 80 percent of households will initially receive only an Internet push mailing, while 20 percent will initially receive a paper form in the mail. Having two different response options from the outset will make it difficult for the Census Bureau to use mass ads and slogans in its social marketing campaign like those used in 2010 ("We can't move forward until you mail it back"). As was true in 2010, the ACS will also be in the field at the same time as the 2020 Census. However, unlike 2010, the ACS will have different data collection operations than the 2020 Census. For example, the initial ACS mailing is Internet push only (forms are only mailed later to nonresponding addresses), and use of a pre-assigned, unique, address-based Census ID is required for online response to the ACS, but not to the 2020 Census. The social marketing campaign for the 2020 Census will need to take both factors into account.

4. Prospects for Reaching Hard-to-Survey Populations in the 2020 Census

The Census Bureau continues to make impressive innovations in the design and implementation of its social marketing campaigns for the decennial census. In her article, Nancy Bates previews several key features of the new campaign for Census 2020, including development of a new self-response propensity for each household, a new survey to understand how the barriers, attitudes, and motivators for households to participate in 2020 have changed since the 2010 Census, and development of a new household-level segmentation system to guide messaging and social marketing plans for hard-to-survey populations.

In planning for 2020, the Census Bureau will also benefit from the ACS – which will continue to provide critical data on computer/device ownership, Internet access, and self-response propensity by mode in the absence of a social marketing campaign. The ACS 2013–2017 five-year estimates for census tracts slated for release in the fall of 2018, will provide a timely update of computer/device ownership, Internet connectivity, and self-response rates by Internet, telephone, and mail for households and small geographic areas across the United States. Of course, technology and Internet access will continue to change before 2020, and the 2013–2017 ACS tract-level estimates will smooth out change across this five-year period. As a result, ACS five-year data may not provide estimates of device

ownership, Internet access, and Internet response in 2018 and 2019 as accurate as those that would be ideal for Census 2020 planning. Nonetheless, ACS data will be a key input for the final operations and social marketing plans for the 2020 Census, including selection of the geographic areas that will initially receive paper forms by mail.

Despite the Census Bureau's innovations and promising social marketing campaign plans, this discussion has highlighted several challenges that may undermine the overall success and accuracy of the 2020 Census. The first is the risk that the undercount of young children will increase in the 2020 Census. While the Census Bureau has conducted an important body of research to better understand the factors associated with this rising undercount, it is not yet clear how these research findings can or will be translated into processes or operations to reduce the undercount in 2020. A related challenge is the fact that changes in family formation processes and living arrangements will continue – and may even accelerate – in the remaining years prior to the 2020 Census. Without clarification of residence rules, improvements in instructions for respondents, and targeted outreach, response rates and data accuracy may be lower in 2020 for the growing share of complex households. The Census Bureau, demographers, and survey researchers all need to evolve our concepts of residency and improve how we measure relationships to more accurately reflect the ways people live together now.

5. References

Baumgardner, S.K., D.H. Griffin, and D.A. Raglin. 2014. "The Effects of Adding an Internet Response Option to the American Community Survey." 2014 American Community Survey Research and Evaluation Report Memorandum Series #ACS14-RER-21, U.S. Census Bureau, Washington, DC. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Baumgardner_04.pdf (accessed October 2017).

Erdman, C. and N. Bates. 2014. "The U.S. Census Bureau Mail Return Rate Challenge: Crowdsourcing to Develop a Hard-to-Count Score." Research Report Series, Statistics #2014-08, U.S. Census Bureau, Washington, DC. Available at: https://www.census.gov/srd/papers/pdf/rrs2014-08.pdf (accessed October 2017).

Manning, W.D. and B. Stykes. 2015. *Twenty-Five Years of Change in Cohabitation in the U.S., 1987–2013*. National Center for Family and Marriage Research Family Profile FP-15-01. Available at: https://www.bgsu.edu/content/dam/BGSU/college-of-arts-and-sciences/NCFMR/documents/FP/FP-15-01-twenty-five-yrs-cohab-us.pdf (accessed October 2017).

Monte, L.M. 2017. "Multiple Partner Fertility Research Brief." *Current Population Reports P70BR-146*. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/publications/2017/demo/p70br-146.pdf (accessed October 2017).

Nichols, E., R. Horwitz, and J.G. Tancreto. 2015. "An Examination of Self-Response for Hard-to-Interview Groups when Offered an Internet Reporting Option for the American Community Survey." 2015 American Community Survey Research and Evaluation Report Memorandum Series #ACS15-RER-10, U.S. Census Bureau, Washington, DC.

Available at: https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Nichols_01.pdf (accessed October 2017).

O'Hare, W.P. 2015. *The Undercount of Young Children in the U.S. Decennial Census*. New York: Springer Briefs in Population Studies.

Ryan, C. and J.M. Lewis. 2017. "Computer and Internet Use in the United States: 2015." *American Community Survey Reports*, ACS-37, U.S. Census Bureau, Washington, DC. Available at: https://www.census.gov/content/dam/Census/library/publications/2017/acs/acs-37.pdf (accessed September 2017).

U.S. Census Bureau. 2014. *The Undercount of Young Children*. Washington, DC. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/demo/2014-undercount-children.pdf (accessed October 2017).

U.S. Census Bureau. 2017a. *Investigating the 2010 Undercount of Young Children – Child Undercount Probes*. 2020 Census Memorandum Series 2017.03. Washington, DC: U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-2017_03-undercount-children-probes.pdf (accessed September 2017).

U.S. Census Bureau. 2017b. *Investigating the 2010 Undercount of Young Children – Analysis of Census Coverage Measurement Results.* 2020 Census Memorandum Series 2017.04. Washington, DC: U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-2017_04-undercount-children-analysis-coverage.pdf (accessed September 2017).

U.S. Census Bureau. 2017c. *Investigating the 2010 Undercount of Young Children – Examining Data Collected during Coverage Followup.* 2020 Census Memorandum Series 2017.05. Washington, DC: U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-2017_05-undercount-children-examining-data.pdf (accessed September 2017).

VanOrman, A. and P. Scommegna. 2016. "Understanding the Dynamics of Family Change in the United States." *Population Bulletin* 71, no. 1. Population Reference Bureau, Washington, DC. Available at: http://www.prb.org/pdf16/prb-population-bulletin-71.1-complex-families-2016.pdf (accessed October 2017).

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 901–919, http://dx.doi.org/10.1515/JOS-2017-0043

Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey

Robert Ashmead[1], *Eric Slud*[1,2], *and Todd Hughes*[3]

The notion of respondent contact burden in sample surveys is defined, and a multi-stage process to develop policies for curtailing nonresponse follow-up is described with the goal of reducing this burden on prospective survey respondents. The method depends on contact history paradata containing information about contact attempts both for respondents and for sampled nonrespondents. By analysis of past data, policies to stop case follow-up based on control variables measured in paradata can be developed by calculating propensities to respond for paradata-defined subgroups of sampled cases. Competing policies can be assessed by comparing outcomes (lost interviews, numbers of contacts, patterns of reluctant participation, or refusal to participate) as if these stopping policies had been followed in past data. Finally, embedded survey experiments may be used to assess contact-burden reduction policies when these are implemented in the field. The multi-stage method described here abstracts the stages followed in a series of research studies aimed at reducing contact burden in the Computer Assisted Telephone Interview (CATI) and Computer Assisted Personal Interview (CAPI) modes of the American Community Survey (ACS), which culminated in implementation of policy changes in the ACS.

Key words: Contact History Instrument; nonresponse follow-up; respondent burden; paradata.

1. Introduction

1.1. General Background on Respondent Contact Burden

It is in the nature of household sample surveys that by responding, a person gives up a certain amount of time, effort, and possibly privacy. Survey researchers refer to this cost as respondent burden or response burden, and they associate increased respondent burden with diminished data quality and response rates (Bradburn 1978; Sharp and Frankel 1983)

[1] U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC, 20233 U.S.A. Emails: eric.v.slud@census.gov, robert.douglas.ashmead@census.gov
[2] Mathematics Department, Kirwan Hall Room 2314, University of Maryland, College Park 20742
[3] UCLA Center for Health Policy Research, 10960 Wilshire Blvd, Suite 1550, Los Angeles, CA 90024, U.S.A. Email: toddhughes@ucla.edu

Acknowledgments: The authors thank Debbie Griffin for initiating and inspiring the work reported here, Asaph Young Chun for critical review and encouraging the publication of this article, and the editors and anonymous referees for helpful comments.

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

both within the given survey and in future surveys. In particular, a respondent who feels burdened by answering one survey might be less likely to accept the next request to participate in a survey. Many authors argue that behavior is primarily affected by the *perceived* respondent burden (Bradburn 1978; Fricker et al. 2014), which filters burden through the respondent's experience. For a recent summary of this literature, see the Eurostat document Hedlin et al. (2005).

In most literature on household surveys, 'respondent burden' refers to the survey instrument or interview process itself (length, difficulty or sensitivity of questions, etc.). In his classic paper on respondent burden, Bradburn (1978) restricted himself to the four general headings, 1) length of interview; 2) amount of effort required by respondent; 3) amount of stress on respondent; and 4) frequency with which respondent is interviewed. Fricker et al. (2014) develop a structural model of respondent burden in terms of three inputs: motivation, task difficulty, and recruitment effort. In this model, difficulty and recruitment are treated as 'related to survey characteristics', with recruitment effort a latent construct with the indicators of *frequent contact*, *frequent personal visit*, and *converted refusal*. The second of these indicators is defined by whether the interviewer visited the respondent more than the median number of visits, so that this and the converted-refusal indicator include survey recruitment preceding the respondent's exposure to the survey instrument. However, in the other elements of the Fricker et al. (2014) model and in almost all other literature not associated with the American Community Survey (ACS), burden was viewed as an aspect of the respondent answering the survey itself, not of the contacts leading to response. Many authors dating back to Bradburn (1978) mentioned the relevance of cumulative burden from all other survey solicitations to a respondent's decision whether or not to participate in a given survey.

In this article, we focus attention on measurement and reduction of what we term the *respondent contact burden* imposed on a potential respondent by the cumulative efforts of survey personnel to make contact with that person or household in any mode, prior to the actual completion of the survey. This can take the form of repeated mailings, telephone calls, personal visits including quick assessments from a car of whether anyone was home or leaving materials at the doorstep, and so on. Especially in surveys where persistent attempts are made to secure interviews in the field, contact burden may be viewed as a social cost that should be mitigated as far as possible without compromising the validity of the survey results. Reduction of contact burden is especially important for large national surveys like ACS which require continuing public acceptance and trust.

An essential part of measuring respondent contact burden and designing nonresponse follow-up strategies is the collection of survey paradata. Paradata might simply be the tallied numbers of visits or calls on each case, but systematic reporting of contact histories by field interviewers has recently become common practice, as formalized in systems such as the Contact History Instrument (Dyer 2004). In lieu of a formal instrument, interviewer observations might be used to summarize or replace paradata (Groves et al. 2007), and characteristics of respondents from historical data (Luiten and Schouten 2013) could serve a related purpose. Groves and Heeringa (2006) are frequently cited as early advocates for the use of paradata in 'responsive' survey design. Many papers (Bates et al. 2010; Maitland et al. 2009; Bates et al. 2008; Olson and Groves 2012) propose to use paradata from contact histories for predicting survey-response propensities. A few authors focus on

level-of-effort paradata (Slud 1998 and 1999; Biemer et al. 2013) and the search for proxy measures (Kreuter et al. 2010) that might predict response rates but not outcomes. Although some of these papers, particularly Bates et al. (2010), emphasize the measurement aspects and data quality of contact histories, there has been little work specifically on the measurement of contact burden with a view to designing interventions to mitigate it.

So-called adaptive interventions or adaptive study designs arise in various branches of statistics, from randomized controlled trials in biomedical statistics (Bothwell and Podolsky 2016) to sample surveys. The adaptive design generally takes the form of some scheduled change in the study protocol as a contingent result of interim outcome data. Because these "interventions" enter *after* the randomization stage, which in the survey context consists in choosing the probability sample, great care must be taken to ensure that the interim outcomes which influence the conduct of the study affect the desired study data as little as possible. In conducting sample surveys, the concern is that altering the data-collection (usually the handling of sampled but not yet responding units at the time of intervention) not affect the quality or values of the data yet to be collected. Data quality, which might be harmed by overly persistent contact strategies, is sometimes assessed within ACS research by the level of item nonresponse, but is generally difficult to measure without an independent audit of the correctness of survey responses.

The most common type of intervention in adaptive study designs is the use of interim outcomes to determine whether follow-up of a not-yet-responding study participant, or the study as a whole, should be terminated. Such interventions are generally rule-based, depending on the values of certain *control variables* up to the interim times when follow-up might be terminated, and the rules are specified as part of the study design before randomization. The control variables governing curtailed follow-up might be summary statistics from the responses collected so far in the entire study, e.g., the total number of sampled survey units who have responded within a geographical area, or might be contact-history variables or model-based predictions of the probability of response, based on case-history data specific to each observational unit whose follow-up might be curtailed. Examples of stopping rules at the whole-survey level include the idea to stop based on levels of the 'R-indicator' as a measure of 'representativeness' (Schouten et al. 2009) or of an estimated probability that a key survey outcome would change more than a specified amount (Wagner and Raghunathan 2010). By contrast, our discussion in this article of adaptive designs supported by contact history paradata is restricted solely to nonresponse-follow-up policies in sample surveys, applied at the level of the individual case.

Our discussion of contact burden concentrates heavily on research done for the ACS. We recognize that ACS is far from typical in the contact burden it imposes. Because of the mandatory nature of the ACS, and the Census Bureau's determination to maintain survey quality through exceptionally high overall response rates (approximately 97% survey-weighted response rate among eligible housing units, over the years 2005–2014), telephone and personal interviewers in the ACS are persistent in their follow-up with sampled housing units. This persistence is extremely successful at securing interviews, even after many prior contact attempts, but may cause persons in sampled units to feel unduly burdened. The research methods developed at the Census Bureau to mitigate this type of burden with minimal adverse impacts, such as lost interviews, are part of a broad effort to reduce the burden experienced by persons sampled in the ACS.

1.2. Multi-Stage Process for Developing and Testing Contact Policies

The theme of this article is that adaptive intervention policies for curtailing burdensome contact follow-up can be developed and tested in a multi-stage process, based on the regular collection of contact history data. Table 1 gives a summary of this multi-stage process (A)–(E) that can be followed closely for repeatedly administered surveys.

The first stage is to identify control parameters (such as maximum number and duration and type of admissible contact attempts) and devise intervention policies in terms of them, subject to administrative or logistical constraints. In a truly adaptive setting, these control parameters would include measures of propensity to respond, developed through models of response and time-dependent covariates fitted to earlier contact history data. Next, relevant outcome measures from existing data sources are specified, such as interview case completion rate and number of contacts and perhaps indicators of reluctance such as refusals by potential respondents, but also some measures of data quality such as item nonresponse or the magnitude of change in key survey estimates from those under the default intervention policy. The third stage consists of a retrospective analysis of contact history data to generate descriptive hypothetical data concerning outcomes that would have been realized had each of the intervention policies been followed. Based on this analysis, a policy is chosen for implementation based on a cost-benefit comparison or trade-off of expected outcome metrics under the policies and on the identification of constraints on allowed values for those outcome measurements. Such constraints would almost always include bounds on the changes in key survey estimates under the changed policies versus the previous policy. The new policies that meet administrative constraints and have admissible cost-benefit profiles are the ones that can actually be chosen. Lastly, the chosen policy is implemented under field conditions that are as close as possible to controlled experimental conditions for comparison with the regime under the earlier policy. Data are collected on contact histories and related outcome measures and assessed for acceptability and possible policy improvements.

The definition of 'admissible' policies requires some clarification. The general meaning of admissibility is that an allowable policy must not be dominated with respect to all costs and benefits by any other feasible policy or probability-weighted combination of policies. In game theory, economics, or statistical decision theory (Ferguson 1967), *multiple* outcome measures (all measured on a scale in which larger signed values are better) together constitute an outcome vector for each feasible policy, and all probability-weighted (i.e., convex) combinations of the feasible outcomes sweep out a convex multidimensional region. The admissible policies are those for which the outcome vector lies on the lower boundary or envelope of this feasible outcome region.

The objective of this article is to show these stages of development of adaptive designs in practice, along with the remaining methodological developments needed to make the design work well, with some focus on limitations of this process as applied so far.

2. ACS Experiments and Analyses on Contact Burden

As part of a broader ACS research effort on the reduction of burden perceived by respondents and potential respondents, the multi-stage process (A)–(E) has been followed

Table 1. Multi-stage development of adaptive policies to reduce burden of nonresponse follow-up.

Stage	Description	Examples
A. Identify control parameters and scope of policies for intervention	Data-collection control parameters that can be manipulated to reduce contact burden	Maximum number, type and duration of contact attempts
B. Specify policies and outcome metrics for evaluating them	Policies controlled and outcomes measured from existing data sources	Interview completion rate, number of contacts, cumulative sum of contacts weighted by burden
C. Analyze retrospective data under intervention policies	Using existing data, calculate outcome metrics as if new policies had been followed	A series of complete survey cycles of contact histories and outcomes
D. Choose a policy for implementation	Compare metric changes under competing policies, identify constraints on outcomes, and choose an admissible policy	Trade-off between number of contacts versus number of lost interviews or other quality measures
E. Assess new policy for acceptable outcomes	Under controlled field conditions, confirm outcome measures fall in desired range, compare operational alternatives	Randomized design elements, in controlled experiment comparing new policy outcomes versus outcomes from earlier period

in research on two separate modes of attempted contact, the Computer Assisted Telephone Interview (CATI) mode and the Computer Assisted Personal Interview (CAPI).

ACS data collection consists of three months of effort for each monthly panel of sampled households (U.S. Census Bureau 2014). In the first month, up to five pieces of mail are sent to sampled households in an effort to obtain an internet or mail self-response. Nonresponse follow-up proceeds to a second month for almost all 'mailable' ACS-sampled housing unit addresses with at least one known telephone number. The CATI operation schedules repeated call attempts terminating either after an interview; after a certain maximum number of call attempts; after reaching control thresholds of numbers of refusals, hang-ups, or unproductive calls; or at the end of the second month. A fraction of about 1/3 of the sampled housing units that do not provide a sufficiently complete CATI interview are sub-sampled into the personal-interviewer CAPI phase of data-collection. In addition, a higher proportion (closer to 2/3) of all national non-mailable sampled addresses are directly sampled into CAPI. The CAPI phase consists of trained Field Representatives (FRs) attempting to secure interviews by personal visits to household or, where possible, by further telephone contact attempts.

Contact history data for CATI are generated through the central calling operations under a system called WebCATI. CAPI field operations are assigned and managed centrally through a Case Management System and tracked and reported through a Unified Tracking System (UTS) based primarily on FR self-reports of case contact attempt histories through the Contact History Instrument (CHI), a centralized online system into which FRs enter details of their work on each case. CHI is described in Dyer (2004) and its design is further discussed by Bates et al. (2010). Although not designed for the purpose of measuring contact burden (Dyer 2004), CHI has been re-designed at least once with that in mind (Virgile 2015). These systems, WebCATI and CHI, may be the only ongoing real-time paradata systems in the U.S. containing longitudinal information on contacts and respondent reluctance for sampled households who choose *not* to respond to a survey.

CATI and CAPI modes – both in ACS and more generally – have unique and distinctly different characteristics. As a result, the control parameters, constraints, potential policies, and metrics that might be used as part of the multi-stage process (A)–(E) may be quite different depending on mode. Still, we argue that the same process can be used in both cases. The ACS CATI and CAPI research on the reduction of respondent burden is documented in a series of reports (Table 2) which collectively represent the multi-stage process (A)–(E). Some of the research documents span multiple stages. In the following sections we summarize the findings from ACS research on CATI and CAPI contact burden. Table 2 is designed to help the reader match the stage of the multi-stage process (A)–(E) to the references.

2.1. *Research on CATI Burden*

In 2012, at the beginning of this line of research, CATI follow-up on a case would stop after the earliest call at which the cumulative total of refusals was at least 2, the total of hang-ups reached 3, the number of 'unproductive' (i.e., unanswered) calls reached 20, or the total number of calls was 25. The Census Bureau hoped to identify strategies that would reduce respondent burden from CATI follow-up with small effects on response

Table 2. Research references on CATI and CAPI burden, by stage.

Stage	CATI	CAPI
A. Identify control parameters and scope of policies	(Zelenak and Davis 2013)	(Zelenak 2014); (Griffin and Nelson 2014)
B. Specify policies and outcome metrics	(Griffin and Hughes 2013); (Slud and Erdman 2013)	(Griffin 2014); (Griffin and Nelson 2014)
C. Analyze retrospective data under interventions	(Griffin and Hughes 2013); (Slud and Erdman 2013)	(Griffin et al. 2015)
D. Choose a policy for implementation	(Griffin and Hughes 2013)	(Griffin et al. 2015)
E. Assess new policy for acceptable outcomes	(Griffin 2013); (Griffin 2014)	(Hughes et al. 2016)

rates. The data analyzed in this research consisted of preprocessed CATI status and history (WebCATI) files and CAPI history (CHI) files covering the national ACS monthly samples for which contacts began during the period June 2011 through Feb. 2012. The data set consisted of 1,097,985 housing-unit records (cases), of which 307,054 were only in CAPI, 600,203 only in CATI, and 190,728 in both.

Zelenak and Davis (2013) summarized ACS CATI case history data, through cross-tabulations of case outcomes (under both CATI and CAPI) with configurations of *control parameters*, by which we mean the numbers of call attempts, of hang-ups, of contacts with members of the sampled household or some other person, and cumulative counts of several categories of expressed reluctance or outright refusal to respond. The main results were tabulations of CATI cases according to the counts of calls, hang-ups, and refusals and the interview status at CATI termination, and the case distributions of final outcome types by total number of calls and of total number of calls by final outcome.

The CATI contact history data were analyzed further in Griffin and Hughes (2013) to suggest specific changes to the rules for termination of CATI case follow-up in terms of the control parameters. It was known to be technologically feasible to manage CATI operations so as to terminate case follow-up when different parameter thresholds were reached. More complicated or contingent interventions in CATI case follow-up were deemed infeasible because they would have required developing and testing new functionality for CATI software. As a result, Griffin and Hughes (2013) proposed 14 (centralized, automatic) termination policies involving reduction of the refusal maximum to 1, the hang-up maximum to 2, the maximum number of unproductive calls to 15 or 12, or the maximum number of calls to 20 or 15. They chose outcome variables from the WebCATI contact histories: total CATI numbers of contact attempts and CAPI workload (the 1/3 of eligible housing-units in CATI not providing CATI interviews that would be subsampled and passed on to CAPI nonresponse follow-up); CATI calls resulting in contacts; a weighted score for calls that combined the cumulative counts of previous refusals, hang-ups, requests to call back; and CATI and CAPI unweighted response rates.

The report then reanalyzed the CATI data according to the outcome metrics that would have been realized under the specified alternative policies.

As a more methodologically oriented complement to the Griffin and Hughes (2013) report, Slud and Erdman (2013) analyzed the same CATI data by calculating the discrete hazard functions associated with interview completion under a time-dependent state measurement defined by Calls, Hang-ups and Refusals. This *discrete hazard function* applied to the case group G at call number k is defined (Klein and Moeschberger 2003) as the proportion of cases in group G followed for k or more calls that terminate in an interview at the k'th call. This function of k may be viewed as the interview yield rate within group G, specific to the number k of calls. As Slud and Erdman (2013) emphasized, this is a conditional response propensity that can be used to suggest adaptive case-follow-up policies for different case groups defined from paradata observable up to each specified number of calls. While Griffin and Hughes (2013, Table 8, p. 10) compared proposed new policies different from the earlier one by calculating ratios of cost saving, calls eliminated, or post-resistance calls eliminated per lost interview, trading off favorable cost and contact outcomes against the unfavorable lost-interview outcomes, Slud and Erdman proposed more complicated and less feasible policies which would extend follow-up only for those groups where the discrete hazard (conditional call-specific interview-yield function) would exceed some threshold. The value of policies based on a decomposition into groups G by contact history depends on verifying that a relatively small set of groups succeeds in strongly separating the group-and-call-specific interview yield. Slud and Erdman (2013, Fig. 1, p. 6) showed that this was indeed possible. The most striking finding was that households that at any stage requested a call back had a strikingly higher propensity to complete an interview, almost regardless of other aspects of their contact history.

As part of the research leading up to the Griffin and Hughes (2013) report, it was confirmed that the ACS centralized call management system could easily be programmed to accommodate changed maximum hang-up and productive- and total- call parameters but not to curtail follow-up based on more complicated combinations of these or other case-specific contact-history variables such as those proposed by Slud and Erdman (2013). Based on the Griffin and Hughes (2013) analysis, the ACS Office determined to change the parameters ending CATI case follow-up beginning with the March 2013 ACS panel (i.e., beginning with CATI operations in April 2013) by reducing the maximum numbers of nonproductive and total calls. Griffin (2013) assessed the changes in experienced outcomes for CATI operations in the months April–May 2013 by comparison with outcomes for the January–February 2013 panels. The field implementation of the CATI call parameters provided at best a very imperfect observational setting, partly because the initial mailings in the January panel collide with the Christmas and New Year holiday period and because January and February tend to have seasonally high mail response rates, but also because new recording technology and telephone scripts were initiated in March 2013 CATI operations. The outcome metrics were observed as part of April and May CATI operations and can be compared either with those from February-March CATI (i.e., those from the January-February panels) or with the ACS June 2011-Feb. 2012 data analyzed in Griffin and Hughes (2013). However, any observed changes could not be ascribed solely to the new CATI call parameter changes. In fact, Griffin (2013) found all of the outcomes (productive and unproductive calls, contacts, interviews, subsampled CAPI

workload) to be in the expected range based on the earlier data analysis, but these comparisons could not be presented with statistical precision due to the uncontrolled (because universal) nature of the implementation.

Griffin and Hughes (2013) broke new ground in quantifying cumulative contact burden by defining a simple *contact score* to summarize a case contact history in terms of severity of burden. Their objective was to provide a single descriptive statistic of burden in terms of which to measure differences between the outcomes of their 14 proposed call-stopping policies versus the previous (up to 2012) call-stopping rules. They displayed these changes in their Table 5, p. 8. This burden score was defined as a case-specific sum over all call attempts, of a score defined at each call attempt as follows. If the case had never resulted in a call-back request at previous calls, the score was 0 if there had been no previous hang-ups or refusals, 1 if there had been one immediate hang-up (but no refusals), 2 if there had been in previous calls either 2 hang-ups or 1 refusal (but not both), and 5 if there had been previously at least one hangup and one refusal. If there had been a previous call-back request by the potential respondent, then the scores under these four different prior-history conditions were defined respectively as 0.5, 0.5, 1, and 3. While these scores were based on introspection by Griffin and Hughes (2013) rather than any research on actual perceived burden by potential respondents, they represent the first example we know of where contact history data are codified into a burden outcome at case level, viewed as a quantity to be minimized.

2.2. *Research on CAPI Burden*

In the past, FRs in ACS relied on their judgment and feedback from their supervisors to determine when to terminate attempted contacts with a sampled case under the CAPI mode of data collection. The ACS Office approved a series of research projects to explore whether contact burden might be lessened by systematic policies for terminating CAPI case follow-up without large losses in response rates. The resulting sequence of research reports, culminating in an August 2015 Pilot study on modified CAPI follow-up control processes, followed the same multi-stage process (A)–(E). First, Zelenak (2014) provided baseline descriptions of CAPI workloads and outcomes from the June 2011–Feb. 2012 dataset previously explored for CATI histories. She began by cross-classifying CATI-nonresponder cases subsampled into CAPI according to their CATI outcomes, and gave descriptive summaries of CAPI workloads and of CAPI cases classified by outcomes and by numbers of contact attempts. From the vantage point of contact burden reduction, the most interesting finding (see Table 8 on page 16 of the report) was the relatively small number of mean and median contact attempts (1 to 3) for cases ending as Interviews or Ineligibles and the much larger (3 to 8) mean and median numbers for cases with ultimate outcomes of Refusals or other Noninterviews.

Next, Griffin and Nelson (2014) created baseline summaries of many different outcomes from the CAPI contact history dataset on January through December 2012 operations. They distinguished response rates from contact rates, cooperation rates and refusal rates, and tabulated outcomes according to whether sampled housing units turned out to be occupied. (This last issue is particular to ACS, where 'vacant' unit interviews related to housing type are collected when FRs can find corroboration of vacancy.) They also measured differences in completeness among interviews based on contact attempts

and levels of cooperation, and delved deeper into strategic patterns followed by FRs, in classifying CAPI workloads according to sequential patterns of telephone and personal visits. Finally, they drew attention to known errors in CHI self-reporting by FRs (Bates et al. 2008) with respect to reluctance displayed by potential respondents.

This research laid the groundwork for the definition of outcome metrics and policy alternatives for curtailing nonresponse follow-up. With the goal of reducing contact burden, Griffin (2014) identified possible changes to CAPI data collection rules, ruling out large proposed system or instrument changes. In addition, it was thought advisable to minimize the impact of proposed changes on the policies and incentives governing FRs. A team of ACS researchers familiar with the CAPI baseline research collaborated in proposing rule changes that might lessen burden without dramatically worsening response rates. Griffin (2014) collected the ideas and distilled them into policy proposals to stop CAPI case follow-up after reaching thresholds of numbers and types of contacts, of estimates of perceived contact burden, or of estimated response propensity. Refinements of these ideas allowing thresholds to depend on CATI outcomes or geography were also considered. This CAPI research generated a list of 27 policies grouped according to which control variables measuring CHI states would trigger termination of CAPI follow-up attempts on a case.

Griffin et al. (2015) analyzed data on the 2012 ACS case histories to tabulate the outcome metrics that would have resulted from implementing each of the 27 CAPI stopping rules. The rules were compared with respect to a large set of different outcomes: contacts, broken down as personal visits or telephone contacts; interviews and non-interviews according to whether or not contact was made with the sampled household; contacts at which either expressed reluctance or 'firm reluctance' was recorded in CHI; contacts after either reluctance or firm reluctance was expressed; and a cumulative burden score for CAPI analogous to the summary contact score previously devised for CATI by Griffin and Hughes (2013). The burden score combined an initial score for the total burden incurred in mail and CATI modes, with an incremental numeric value from 1 to 15 for each CAPI contact attempt, based on an intuitive scoring of the relative burden of various types of contact coded in CHI. In this scoring, personal visits were assigned more burden points than telephone calls, contacts more than non-contacts, and additional points were added if a potential respondent indicated reluctance. The cumulative burden score served as the primary proxy for respondent perceived contact burden. All of the different outcomes analyzed were alternative measurements for workload/cost or nonresponse or perceived contact burden, and the approach was to balance increases in nonresponse due to implementing the new stopping policies versus reductions in contacts or contact burden. This balancing was done in the same sequence as for CATI policy changes, first through cost-benefit ratios (Table 15 of Griffin et al. 2015) measuring the reduction in cost or contacts per lost interview, or reduction in burden-score or highly-burdened cases per lost interview, and then (Figures 1–2 of the same report) through the nearness of the vector of three policy outcomes for each policy – mean contacts, mean burden score, and case non-interview rate – to the lower boundary of the convex combinations of all such policy outcome vectors. This lower boundary is understood as the result of cost-benefit trade-offs, providing the set of best policy outcomes that can be achieved by applying the different policies with various probabilities. A small set of competing policies were thus identified as most interesting, further reduced to a single recommended policy after clarifying the administrative

constraint that an overall CAPI nonresponse rate greater than about 0.085 could not be borne. The chosen policy, number 18 among those displayed, was a stopping rule to terminate CAPI contacts on cases that exceed the threshold of 40, a level roughly corresponding to the 90'th percentile of cumulative burden among all CAPI cases in 2012 ACS data.

With respect to lost interviews and reduced burden, the stopping policy chosen for implementation was the most aggressive of those near the lower envelope of the nonresponse, contact and burden score outcome combinations that met the upper-bound constraint of 0.085 on CAPI nonresponse (Griffin et al. 2015). Interestingly, the policy chosen was not one of the three competing policies defined in terms of thresholded response propensities, but rather was designed to achieve maximum reductions in the burden score. It resembled the CATI stopping-policy chosen for implementation in 2013 through its explicit definition in terms of parameters controlling contact burden. The avowed purpose of both research initiatives had been to find a stopping policy to achieve maximum reduction of contact burden subject to small and acceptable losses in interview completion rates. The CAPI policy of stopping according to a burden-score threshold was chosen for actual (pilot) implementation only after additional research (not published) verifying on the ACS 2012 data that this policy did not lead to unacceptably large CAPI nonresponse or small contact and burden reductions when tabulated by state and Survey Statistician Field Area (a set of 48 geographic organizational units for CAPI field work).

In order to test the effectiveness of the cumulative burden score stopping rule under field conditions, the Census Bureau conducted a field pilot in the CAPI operation of the ACS during August 2015 (Hughes et al. 2016). The pilot required that FRs transmit data from their laptops to a central management system (UTS) twice daily on days worked, at the beginning of the work-day before attempting contacts and after all of the day's attempts, so that burden scores could be updated correctly. The transmission prior to contact attempts updated the case burden scores with information from the day before and removed any cases with burden scores above the burden threshold from FRs' laptops. A secondary goal of the pilot was to learn whether survey outcomes would be affected by displaying current burden scores on FR laptops. The previous ACS protocol required FRs to transmit only once per day, at the end of their workday. The UTS system updated the cumulative case burden scores using entries from the FR in CHI. Therefore, the cumulative burden score calculation depended on accurate and timely FR entry of case-specific data. The CHI includes information on each contact attempt made to sample units as well as additional levels of effort not classified as contact attempts (e.g., locating the unit or geocoding).

The pilot was conducted in one-quarter of the field geographies for ACS interviewing. The remaining field geographies followed the standard field protocol. Within the Survey Statistician Field Areas (SSFAs) selected for inclusion in the pilot, individual Field Supervisor (FS) areas and all the FRs within that FS area (FSA) were assigned randomly to one of three experimental treatments:

- Treatment 1 (Control): burden score not displayed, cases not removed;
- Treatment 2: burden score displayed, cases removed; and
- Treatment 3: burden score not displayed, cases removed.

Overall, 4.5 percent and 4.1 percent of cases were pulled from Treatments 2 and 3 respectively as a consequence of reaching the cumulative burden score threshold. Only

Table 3. 2016 CAPI pilot estimated differences in outcomes per case.

	Trt 1	Trt 2–3	Difference	Percent difference	*p*-value
Contact attempts	3.90	3.67	0.234	6%	0.074
Contacts	1.00	0.94	0.062	6%	0.067
Contacts with firm reluctance	0.08	0.07	0.017	21%	0.032
Response rate	93.1%	91.8%	1.3%	1.4%	0.104

Source: American Community Survey paradata, 2015.

minimal differences between Treatment 2 and Treatment 3 were observed, and therefore Treatment 2 and 3 observations were pooled in order to increase the precision of estimates of differences with Treatment 1. Table 3 shows the 6% decrease in both contact attempts and contacts per case observed for Treatments 2–3 compared with Treatment 1, along with a 21% reduction in contacts with firm reluctance per case. The CAPI response rate for Treatment 1 was 93.1% compared to 91.8% for Treatments 2–3. The estimated difference was 1.3 percentage points [p-value = 0.104]. The estimated difference 1.3% was relatively small compared to the approximately 4% of cases that were stopped due to high burden score. This suggests that a large proportion of the stopped cases would not have resulted in complete interviews had FRs continued to make attempts.

In the pilot, the cumulative burden score stopping rule succeeded in reducing the percentage of cases with extremely high burden score. Treatments 2–3 had only 0.3 percent of cases with burden scores over 60 (and less than 0.1 percent with scores over 80), while control and Treatment 1 had over 2.0 percent of cases with burden scores over 60 (and 0.5 percent of cases with burden scores over 80). Treatments 2–3 had cases with burden scores as high as 60 or 80 only because of failures to comply with the beginning- and end-of-workday transmission protocol updating the burden score, together with multiple intra-day contact attempts by some FRs.

As a result of the Hughes et al. (2016) report, the Census Bureau decided to implement the cumulative burden score stopping rule nationwide beginning in June 2016. As adopted, the protocol displayed a burden category to FRs, but not the exact score.

2.3. *Before-and-After Assessment*

At the end of the process of Table 1 for contact burden reduction, what can be said about the overall actual impact of the changes? In the ACS CATI and CAPI examples discussed at length above, the final step E was an assessment based on newly collected ACS data, and this can be enriched in each example with the retrospective data-analysis results done in step C to inform the choice of a contact-burden reduction policy.

As mentioned in the ACS reports cited in the paragraphs above, statistical inferences for the effects of the policy changes are generally not available because of the need to correct for period and regime differences, which could only be done through a model. For example, Table 4 shows the per-case number of calls (contact attempts) and CATI interview completion rate for CATI cases, based on the June 2011 – Feb. 2012 data analyzed in Zelenak and Davis (2013), Griffin and Hughes (2013), and Slud and Erdman (2013), and on the comparison in Griffin (2013) of CATI outcomes just before and after

Table 4. CATI call and interview rates per case, before and after rule change.

Stopping rule	Time period	Attempts/case	Interviews/case
Old	June 2011-Feb. 2012	7.25	0.242
Old	Jan.-Feb. 2013	6.70	0.168
New*	June 2011-Feb. 2012	5.72	0.215
New	Mar.-Apr. 2013	5.16	0.150

Source: Zelenak and Davis (2013, Table 4); Griffin and Hughes (2013, Tables 1,4)
*Outcomes calculated for New rule as though followed in the older data set.

the implementation of the new stopping rule in March 2013 with reduced maximum numbers of nonproductive and total calls. Because of a decreasing trend over the past several years in CATI interview-completion rates, the attempt and interview rates are very different across the two time periods. However, the reductions in outcome measures are similar for both periods, in the range 21–23% for attempts per case and roughly 11% for reductions in interview completion rates.

In the progression to a new policy for terminating nonresponse follow-up by FRs in CAPI, Table 5 similarly shows performance through the August 2015 ACS pilot study described above and in Hughes et al. (2016). As in Table 4, the comparison is partly observational, with results displayed for the month before the pilot, July 2015, and for the "Treatment Group 0" which we use to denote the control group of ACS data collections in SSFAs not chosen for inclusion in the pilot. The SSFA choice was not random, and contrasts with that control group and with the ACS outcomes in the month before the pilot do not yield statistical inferences, only descriptive comparisons. The designation of treatment group in data collected in July 2015 arises because data from the Field Supervisory Units (FSAs) that were assigned to the separate treatments in the August pilots are tallied separately in July. The numbers of attempts per case are systematically different in 2015 from their level in 2012, a change that can largely be ascribed to a redesign of the CHI system implemented in January 2014 that explicitly aimed to reduce the potential for underreporting contact attempts. The overall pattern of Table 4 shows a

Table 5. CAPI attempt, interview and burden per case, before and after rule change.

Rule/regime	Time period	Attempts/case	Interviews/case	VHBurd rate
Old	all 2012	2.94	0.950	0.050
Trtgp 0	July 2015	3.84	0.931	0.062
Trtgp 1	July 2015	3.89	0.931	0.060
Trtgp 2–3	July 2015	3.92	0.936	0.062
Trtgp 0	Aug. 2015	3.87	0.934	0.067
Trtgp 1	Aug. 2015	3.90	0.935	0.063
New*	all-2012	2.79	0.920	0.019
Trtgp 2–3	Aug. 2015	3.67	0.918	0.041

Source: Hughes et al. (2016, Tables 5.7, 5.15) and ACS paradata, 2015.
VHBurd is the case indicator of having a burden score above 45.
*Outcomes calculated for New rule as though followed in the older data set.

clear and similar decrease in attempts per case for the new versus old stopping regime, for each separate year of data collection. There was a clear decrease of roughly 2% in CAPI response rate, and a dramatic decrease in the incidence of cases with 'very high' burden score (i.e., above 45, which was the 95'th percentile level for burden score in the ACS 2012 paradata). The decrease in VHBurd from implementation of the new policy was very different in the 2012 data and the 2015 pilot because cases exceeding the threshold 40 defining the contact-attempt stopping rule could not in practice be put into effect instantly but waited for a further transmission from the FR, allowing further contact with the same case during the same day (or the next, in case of transmission delay). Thus, the VHBurd reduction achieved was much less than the possible reduction if case follow-up had been stopped immediately.

The contrast between treatment groups within the August 2015 pilot was based on randomization and therefore did yield meaningful variances and confidence intervals for contrasts between the Treatment group and the combination of Treatment groups 2–3. As described in Hughes et al. (2015), the reduction in attempts per case and CAPI response rate between Treatment Group 1 (which did not remove cases from the FR's case portfolio) and Groups 2–3 (which did) were found to be nearly significant with p-values .074 and .104. A similar permutational calculation for VHBurd reduction in the August pilot yields a highly significant result (p-value <0.001).

3. Limitations of Multistage Approach to Burden Reduction

The contact history databases we are aware of require FRs to enter paradata accurately. However, FRs may have competing incentives not to do so. FRs are evaluated (in part) on their completion rate, and in the context of the ACS CAPI pilot study, when a case is pulled in the future from a FR's workload for exceeding the burden score, the case may count against that FR's response rate. This may influence FRs not to record some contact attempts that result in high amounts of respondent contact burden. In the pilot study, there was little evidence that such behavior occurred; however, the removed cases did not count against the FRs during that month, and FRs were found generally not to have a good understanding of the burden score increments, knowledge that they might acquire over time. This raises the importance of heightened monitoring of field reporting behaviors after the full implementation of the new stopping rule. Additionally, the chosen measures of contact burden in studies conducted up to the present, such as the burden score increments used in the ACS CAPI stopping rules based on cumulative burden score, have never been validated against actual perceptions of contact burden by potential respondents. Cognitive research along this line is needed, specifically to validate the current score used in ACS or to suggest modifications.

Even when contact history data have been properly recorded and maintained, retrospective analysis of those data may not be a reliable guide to factors influencing response rates and other measures of survey quality when new procedures are implemented. For this reason, the effects of implementation should always be tested in the field, ideally in a randomized study permitting assessment of the relative effects of the new policy versus other less predictable changes in the survey environment. That is the motivation for including stage E in Table 1.

Conducting randomized controlled experiments in a survey field setting can be difficult in part because of compliance failures and administrative procedures which either allow participants to know how their treatment group differs from others or which change the treatment for some participants after randomization. Analogous issues arise in randomized clinical studies where patients do not fully comply with treatment regimens, for example by failing to take pills on schedule, or where they or their doctors can infer their treatment assignment, or where there is some unanticipated treatment switching. All of these failures to follow the formal randomization protocol did occur in the 2015 ACS CAPI pilot of the cumulative burden score stopping rule. Among the limitations in the pilot were the reassignment of cases to FRs in different treatment groups, and transmission compliance was much lower than was needed to ensure that burden scores were daily entered accurately and cases were pulled in a timely manner. In addition, because cases were rarely supposed to reach the burden score threshold, the implementation of the stopping rule was not expected to cause drastic changes in metrics of contact or response. For these reasons, and also because the geographies chosen for the pilot were not a random sample of all ACS geographies, the results were not representative of the whole ACS. Thus the pilot study was partly observational rather than controlled.

4. Conclusions and Directions for Research

In this article, we proposed a multi-stage process for the development of adaptive policies to reduce the burden of nonresponse follow-up and described how it has been implemented in two response modes in the ACS. We argued that in a repeatedly administered survey like the ACS, which collects contact history paradata, this multi-stage process can be replicated to create successful burden reduction policies. We also described the cumulative burden scores that have been used by the ACS to study the contact-burden reduction of newly implemented CATI and CAPI contact-attempt stopping rules. In CAPI, a stopping rule for contact attempts based on reaching a certain threshold in the cumulative burden score was chosen for a pilot study and national implementation. The idea of burden scoring as a sum of increments with severity determined by contact history paradata is one that could be adopted and modified for other surveys as a method to measure contact burden. For survey organizations interested in monitoring and mitigating contact burden on potential respondents, it is essential first to develop a system to collect paradata concerning contact attempts. Then, after a period of quantifying baseline measures of contact burden and other process outcomes, the multi-stage process in Table 1 can be undertaken.

Additional iterations of the multi-stage process can be beneficial as conditions and resources change. As a sequel to the research described in Section 2.1, Mills (2016) analyzed CATI contact histories during the six-month period following a change in 2016 of telephone number sources from third-party vendors to the Census Bureau's Center for Administrative Records Research and Applications. Motivated by the dual goals of increasing the efficiency of CATI and decreasing respondent burden, Mills used CATI data to simulate policies altering the control parameters of total-workload, maximum-follow-up, and maximum-follow-up after contact. This work – analyzing new CATI data, choosing outcome measures reflecting cost and efficiency of producing completed

interviews in CATI, and proposing policies reducing case follow-up based on model-based scores for reliability of telephone numbers or on time to first contact – cuts across the stages (A)–(C) of Table 1.

In the experience of Hughes et al. (2016), the assessment under field conditions of implemented adaptive policies to curtail nonresponse follow-up required detailed attention to the impact on morale and incentives of field staff. Since the paradata on which stopping policies are based depend until now on FR self-reporting, the lack of research on the interaction between FR incentives and contact burden is a limitation on the effectiveness of field experiments. More broadly, either because of FR behavioral changes in adapting to changes in their compensation and incentives, or because policy changes induce changes in respondent propensities over time, conclusions from analysis of contact histories must be confirmed or re-evaluated after adaptive policies are implemented.

If survey organizations are seriously concerned about respondent contact burden, then the instruments used to measure it, such as CHI, should be redesigned. Additionally, policies should be evaluated from the perspective of the potential respondent. Though policies such as the cumulative burden score stopping rule are clearly successful at reducing burden according to its own metric, it is not yet clear what the effect is on the potential respondent's perception of burden.

5. References

Bates, N., J. Dahlhamer, P. Phipps, A. Safir, and L. Tan. 2010. "Assessing Contact History Paradata Quality Across Several Federal Surveys." In JSM Proceedings, Survey Research Methods Section, American Statistical Association, Vancouver, BC, July 31-August 5, 2010. Alexandria, VA: American Statistical Association. 91–105. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/y2010/Files/306005_55654.pdf (accessed February 2017).

Bates, N., J. Dahlhamer, and E. Singer. 2008. "Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse." *Journal of Official Statistics* 24: 591–612.

Biemer, P.P., P. Chen, and K. Wang. 2013. "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society:* Series A 176: 147–168. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01058.x.

Bothwell, L. and S. Podolsky. 2016. "The Emergence of the Randomized, Controlled Trial." *New England Journal of Medicine* 375: 501–504. Doi http://dx.doi.org/10.1056/NEJMp1604635.

Bradburn, N. 1978. "Respondent Burden." In JSM Proceedings, Survey Research Methods Section, American Statistical Association, San Diego, California, August 14–17, 1978. Alexandria, VA: American Statistical Association. 35–40. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1978_007.pdf (accessed February 2017).

U.S. Census Bureau. 2014. Design and Methodology: American Community Survey. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html (accessed February 2017).

Dyer, W. 2004. "Contact History Instrument." In IBUC Proceedings, International Blaise Users Conference, Qubec, Canada, September 22–24, 2004. International Blaise Users Group. 35–61. Available at: www.blaiseusers.org/2004/papers/03.pdf (accessed February 2017).

Ferguson, T. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press.

Fricker, S., T. Yan, and S. Tsai. 2014. "Response Burden: What Predicts It and Who is Burdened Out?" In AAPOR Proceedings, American Association for Public Opinion Research, Anaheim, California, May 15–18, 2014. Oakbrook Terrace, IL: American Association for Public Opinion Research. 4568–4577. Available at: https://www.bls.gov/osmr/pdf/st140170.pdf (accessed February 2017).

Griffin, D. 2013. "Effect of Changing Call Parameters in the American Community Survey's Computer Assisted Telephone Interviewing Operation." American Community Survey Research and Evaluation Report Memorandum Series ACS13-RER-17. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Griffin_03.pdf (accessed February 2017).

Griffin, D. 2014. "Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 2 Results." American Community Survey Research and Evaluation Report Memorandum Series ACS14-RER-07. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Griffin_01.pdf (accessed February 2017).

Griffin, D. and T. Hughes. 2013. "Analysis of Alternative Call Parameters in the American Community Survey's Computer Assisted Telephone Interviewing." American Community Survey Research and Evaluation Report Memorandum Series ACS13-RER-11. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Griffin_02.pdf (accessed February 2017).

Griffin, D. and D. Nelson. 2014. "Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 1 Results (Part 2)." American Community Survey Research and Evaluation Report Memorandum Series ACS14-RER-22. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Griffin_02.pdf (accessed February 2017).

Griffin, D., E. Slud, and C. Erdman. 2015. "Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operations – Phase 3 Results." American Community Survey Research and Evaluation Report Memorandum Series ACS14-RER-28-R1. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Griffin_01.pdf (accessed February 2017).

Groves, R. and S. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistics Society: Series A* 169: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Groves, R.M., J. Wagner, and E. Peytcheva. 2007. "Use of Interviewer Judgments about Attributes of Selected Respondents in Post-Survey Adjustment for Unit Nonresponse: An Illustration with the National Survey of Family Growth." In JSM Proceedings, Survey Research Methods Section, American Statistical Association, Salt Lake City, Utah, July 29-August 2, 2007. Alexandria, VA: American Statistical Association. 3428–3431. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/y2007/Files/JSM2007-000782.pdf (accessed February 2017).

Hedlin, D., Dale, T., Haraldsen, G., and Jones, J. eds. 2005. *Developing Methods for Assessing Perceived Response Burden*. Research Report. Stockholm: Statistics Sweden, Oslo: Statistics Norway, and London: Office for National Statistics. Available at: http://ec.europa.eu/eurostat/documents/64157/4374310/10-DEVELOPING-METHODS-FOR-ASSESSING-PERCEIVED-RESPONSE-BURDEN.pdf/1900efc8-1a07-4482-b3c9-be88ee71df3b (accessed February 2017).

Hughes, T., E. Slud, R. Ashmead, and R. Walsh. 2016. "Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation." American Community Survey Research and Evaluation Report Memorandum Series #ACS16-RER-07. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Hughes_01.pdf (accessed February 2017).

Klein, J. and M. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. Springer-Verlag.

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

Luiten, A. and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society:* Series A 176: 169–189. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01080.x.

Maitland, A., C. Casas-Cordero, and F. Kreuter. 2009. "An Evaluation of Nonresponse Bias using Paradata from a Health Survey." In JSM Proceedings, Survey Research Methods Section, American Statistical Association, Washington, DC, August 1–6, 2009. Alexandria, VA: American Statistical Association. 370–378. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/y2009/Files/303004.pdf (accessed February 2017).

Mills, G. 2016. "Simulated Effects of Changing Calling Parameters and Workload Size on Computer Assisted Telephone Interview Productivity in the American Community Survey." American Community Survey Research and Evaluation Report Memorandum Series #ACS16-RER-22. Washington, DC: U.S. Census Bureau. Available at: https://census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Mills_02.pdf (accessed August 2017).

Olson, K. and R. Groves. 2012. "An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period." *Journal of Official Statistics* 28: 29–51.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

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

Slud, E. 1998. "Predictive Models for Decennial Census Household Response." In JSM Proceedings, Survey Research Methods Section, American Statistical Association, Dallas, Texas, August 9–13, 1998. Alexandria, VA: American Statistical Association. 272–277. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1998_043.pdf (accessed February 2017).

Slud, E. 1999. "Analysis of 1990 Decennial Census Checkin-Time Data." In FCSM Proceedings Federal Committee on Statistical Methodology, Washington, DC, November 15–17, 1999. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/IX-B_Slud_FCSM1999.pdf (accessed August 2017).

Slud, E. and C. Erdman. 2013. "Adaptive Curtailment of Survey Follow-up Based on Contact History Data." In FCSM Proceedings Federal Committee on Statistical Methodology, Washington, DC, November 4–6, 2013. 1–11. Available at: https://s3.amazonaws.com/sitesusa/wpcontent/uploads/sites/242/2014/05/B1_Slud_2013FCSM.pdf (accessed August 2017).

Virgile, M. 2015. "Measurement Error in American Community Survey Paradata and 2014 Redesign of the Contact History Instrument." Center for Statistical Research and Methodology Report Series (Survey Methodology #RSM2016-01). Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/srd/papers/pdf/RSM2016-01.pdf (accessed February 2017).

Wagner, J. and T. Raghunathan. 2010. "A New Stopping Rule for Surveys." *Statistics in Medicine* 29: 1014–1024. Doi: http://dx.doi.org/10.1002/sim.3834.

Zelenak, M.F. and M. Davis. 2013. "Impact of Multiple Contacts by Computer-Assisted Telephone Interview and Computer-Assisted Personal Interview on Final Interview Outcome in the American Community Survey." American Community Survey Research and Evaluation Report Memorandum Series #ACS13-RER-08. Washington, DC: U.S. Census Bureau. Available at: www.census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Zelenak_01.pdf (accessed February 2017).

Zelenak, M.F. 2014. "Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 1 Results (Part 1)." American Community Survey Research and Evaluation Report Memorandum Series #ACS14-RER-06. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Zelenak_01.pdf (accessed February 2017).

Received July 2016
Revised February 2017
Accepted September 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 921–962, http://dx.doi.org/10.1515/JOS-2017-0044

Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC)

Laura Boeschoten[1], *Daniel Oberski*[2], *and Ton de Waal*[3]

Both registers and surveys can contain classification errors. These errors can be estimated by making use of a composite data set. We propose a new method based on latent class modelling to estimate the number of classification errors across several sources while taking into account impossible combinations with scores on other variables. Furthermore, the latent class model, by multiply imputing a new variable, enhances the quality of statistics based on the composite data set. The performance of this method is investigated by a simulation study, which shows that whether or not the method can be applied depends on the entropy R^2 of the latent class model and the type of analysis a researcher is planning to do. Finally, the method is applied to public data from Statistics Netherlands.

1. Introduction

National Statistical Institutes (NSIs) often use large data sets to estimate population tables covering many different aspects of society. One way to create these rich data sets as efficiently and cost effectively as possible is to utilize already available register data. This has several advantages. First, known information is not collected again by means of a survey, saving collection and processing costs, as well as reducing the burden on the respondents. Second, registers often contain very specific information that could not have been collected by surveys (Zhang 2012). Third, statistical figures can be published more quickly, as conducting surveys can be time consuming. However, when more information is required than is already available, registers can be supplemented with survey data (De Waal 2016). Caution is then advised, as surveys likely contain classification errors. When a data set is constructed by integrating information at micro-level from both registers and surveys, we call this a composite data set. More information

[1] Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, PO Box 90153, Tilburg 5000 LE, Netherlands and Centraal Bureau voor de Statistiek – Process development and methodology Henri Faasdreef 312, Den Haag 2492 JP, The Netherlands. Email: l.boeschoten@tilburguniversity.edu
[2] Universiteit Utrecht – Social and Behavioural Sciences, Utrecht, Utrecht, The Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, Tilburg, The Netherlands. Email: d.l.oberski@uu.nl
[3] Centraal Bureau voor de Statistiek – Process development and methodology Den Haag, The Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, Tilburg, The Netherlands. Email: T.deWaal@cbs.nl
Acknowledgments: The authors would like to thank the associate editor and the reviewers for their useful comments. Furthermore, the authors would like to thank Barry Schouten and Frank Bais for providing us with the application data.
Simulation code can be found on https://github.com/lauraboeschoten/MILC

on how to construct such a composite data set can be found in Zhang (2012) and Bakker (2010). Composite data sets are used by, among others, the Innovation Panel (Understanding Society 2016), the Millennium Cohort Study (UCL Institute of Education 2007), the Avon Longitudinal Study of Parents and Children (Ness 2004), the System of Social Statistical Databases of Statistics Netherlands, and the 2011 Dutch Census (Schulte Nordholt et al. 2014).

When using registers for research, we should be aware that they are collected for administrative purposes so they may not align conceptually with the target and can contain process delivered classification errors. These may be due to mistakes made when entering the data, delays in adding data to the register (Bakker 2009) or differences between the variables being measured in the register and the variable of interest (Groen 2012). This means that both registers and surveys may contain classification errors, although originating from different types of sources. This assumption is in contrast to what many researchers assume, namely that either registers or surveys are error-free. To illustrate, Schrijvers et al. (1994) used registers to validate a postal survey on cancer prevalence, Turner et al. (1997) used Medicare claims data to validate a survey on health status, and Van der Vaart and Glasner (2007) used optician database information to validate a telephone survey. In contrast, Jörgren et al. (2010) used a survey to validate the Swedish rectal cancer registry and Robertsson et al. (1999) used a postal survey to validate the Swedish knee arthroplasty register. Since neither surveys or registers are free of error, it is most realistic to approach them both as such. Therefore, we aim to develop a method which incorporates information from both to estimate the true value, without assuming that either one of them is error-free.

To distinguish between two types of classification errors, we classify them as either visibly or invisibly present. Both types can be estimated by making use of new information that is provided by the composite data set. Invisibly present errors in surveys or registers can be detected when responses on both are compared in the composite data set. Differences between the responses indicate that there is an error in one (or more) of the sources, although it is at this point unclear which score(s) exactly contain(s) error. The name 'invisibly present errors' is given because these errors could not have been seen in a single data set. They can be dealt with by estimating a new value using a latent variable model. To estimate these invisibly present errors using a latent variable model, multiple indicators from different sources within the composite data that measure the same attribute are used. This approach has previously been applied using structural equation models (Bakker 2012; Scholtus and Bakker 2013), latent class models (Biemer 2011; Guarnera and Varriale 2016; Oberski 2015) and latent markov models (Pavlopoulos and Vermunt 2015). Latent variable models are typically used in another context, namely as a tool for analysing multivariate response data (Vermunt and Magidson 2004).

Covariates (variables within the composite data set that measure something other than the attribute of interest) can help improve the latent variable model. Some errors can then be observed already when an impossible combination between a score on the attribute and a covariate is detected, which we define as a visibly present error. The name 'visibly present errors' is given here because (some of) these errors are visible in a single data set. An example of a combination which is not allowed is the score "own" on the variable *home ownership* and the score "yes" on the variable *rent benefit*. Such an, in practice,

impossible combination can be replaced by a combination that is deemed possible. Whether a combination of scores is possible and therefore "allowed" is commonly listed in a set of edit rules. An incorrect combination of values can be replaced by a combination that adheres to the edit rules. Different types of methods are used to find an optimal solution for different types of errors (De Waal et al. 2012). For errors caused by typing, signs or rounding, deductive methods have been developed by Scholtus (2009, 2011). For random errors, optimization solutions have been developed such as the Fellegi-Holt method for categorical data, the branch-and-bound algorithm, the adjusted branch-and-bound algorithm, nearest-neighbour imputation (De Waal et al. 2011, 115–156) and the minimum adjustment approach (Zhang and Pannekoek 2015). Furthermore, imputation solutions, such as nonparametric Bayesian multiple imputation (Si and Reiter 2013) and a series of imputation methods discussed by Tempelman (2007) can be used.

The solutions discussed two paragraphs above for invisibly present errors are not tailored to handle the invisibly and visibly present errors simultaneously, and they do not offer possibilities to take the errors into account in further statistical analyses; they only give an indication of the extent of the classification errors. In addition, uncertainty caused by both visibly and invisibly present errors is not taken into account when further statistical analyses are performed. An exception is the method developed by Kim et al. (2015), which simultaneously handles invisibly and visibly present errors using a mixture model in combination with edit rules for continuous data, and which has been extended by Manrique-Vallier and Reiter (2016) for categorical data. This method allows for an arbitrary number of invisible errors based on one file and one measurement, whereas we consider multiple linked files with multiple measurements of an attribute. Any method dealing with visibly or invisibly present classification errors should account for the uncertainty created by these errors. This can be done by making use of multiple imputations (Rubin 1987), and has previously been used in combination with solutions for invisibly present errors (Vermunt et al. 2008) and visibly present errors (Si and Reiter 2013; Manrique-Vallier and Reiter 2013).

We propose a new method that simultaneously handles the three issues discussed: it handles both visibly and invisibly present classification errors and it incorporates them both, as well as the uncertainty created by them, when performing further statistical analysis. By comparing responses on indicators measuring the same attribute in a composite data set we allow the estimation of the number of invisibly present errors using a Latent Class (LC) model. Visibly present errors are handled by making use of relevant covariate information and imposing restrictions on the LC model. In the hypothetical cross table between the attribute of interest and the restriction covariate, the cells containing a combination that is in practice impossible are restricted to contain zero observations. These restrictions are imposed directly when the LC model is specified. To also take uncertainty created by the invisibly and visibly present errors into account when performing further statistical analyses, we make use of Multiple Imputation (MI). Because MI and LC are combined in this new method, the method will be further denoted as MILC.

In the following section, we describe the MILC method in more detail. In the third section, a simulation study is performed to assess the novel method. In the fourth section, we apply the MILC method on a composite data set from Statistics Netherlands.

2. The MILC Method

The MILC method takes visibly and invisibly present errors into account by combining Multiple Imputation (MI) and Latent Class (LC) analysis. Figure 1 gives a graphical overview of this procedure. The method starts with the original composite data set comprising L measures of the same attribute of interest. In the first step, m bootstrap samples are taken from the original data set. In the second step, an LC model is estimated for every bootstrap sample. In the third step, m new empty variables are created in the original data set. The m empty variables are imputed using the corresponding m LC models. In the fourth step, estimates of interest are obtained from the m variables and in the last step, the estimates are pooled using Rubin's rules for pooling (Rubin 1987, 76). These five steps are now discussed in more detail.

The MILC method starts by taking m bootstrap samples from the original composite data set. These bootstrap samples are drawn because we want the imputations we create in a later step to take parameter uncertainty into account. Therefore, we do not use one LC model based on one data set, but we use m LC models based on m bootstrap samples of the original data set (Van der Palm et al. 2016).

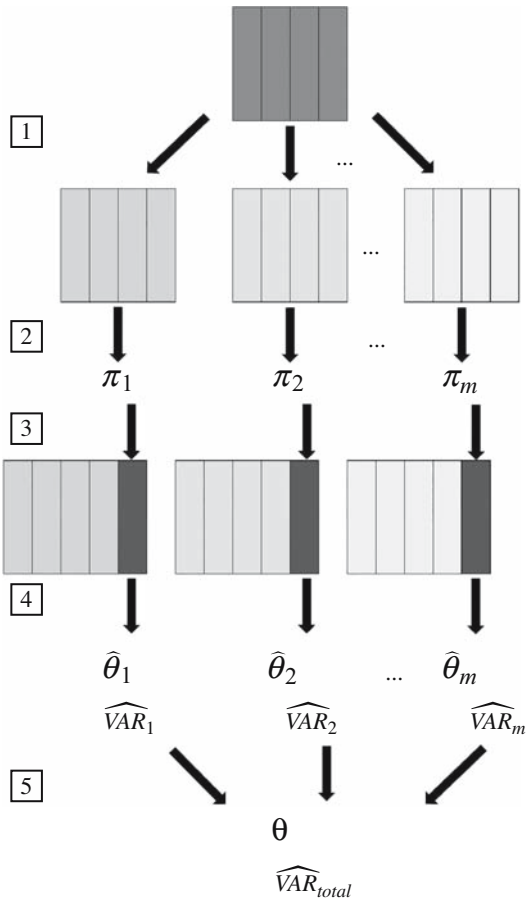

Fig. 1. Procedure of latent class multiple imputation for a multiply observed variable in a composite data set.

In the next step, we make use of LC analysis to estimate both visibly and invisibly present classification errors in categorical variables. We first link several data sets by unit identifiers, resulting in a composite data set matched on a common core set of identifiers (discarding all records where no match is obtained), and group variables measuring the same attribute present on more than one of the original source data sets. For each of the variable groups, we build a single latent variable (denoted by X) representing the underlining true measure, assuming discrepancies between different sourced measures.

For example, we have L dichotomous indicator variables ($Y_1, \ldots, Y_L$) measuring the same attribute *home ownership* (1 = "own", 2 = "rent") in multiple data sets linked on unit level. Differences between the responses of a unit are caused by what we described as invisibly present classification error in one (or more) of the indicators. Since the indicators all have an equal number of categories (C), we fix the number of categories of the latent variable X to C.

The LC model we then build using the indicator variables is based on five assumptions. The first assumption pertains to the marginal response pattern $\mathbf{y}$, which is a vector of the responses to the given indicators. For example, we have three indicators measuring home ownership, the response pattern $\mathbf{y}$ can be "own", "own", "rent". We assume here that the probability of obtaining this specific marginal response pattern $P(\mathbf{Y} = \mathbf{y})$ is a weighted average of the X class specific probabilities $P(\mathbf{Y} = \mathbf{y}|X = x)$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x). \tag{1}$$

Here, $P(X = x)$ denotes the proportion of units belonging to category x in the underlying true measure, where x might be "own", the proportion of the population owning their own house.

The second assumption is that the observed indicators are independent of each other given a unit's score on the underlying true measure. This means that when a mistake is made when filling in a specific question in a survey, this is unrelated to what is filled in for the same question in another survey or register. This is called the assumption of local independence,

$$P(\mathbf{Y} = \mathbf{y}|X = x) = \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{2}$$

Combining Equation (1) and Equation (2) yields the following model for response pattern $P(\mathbf{Y} = \mathbf{y})$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)\prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{3}$$

The model parameters ($P(X = x)$ and $P(Y_l = y_l|X = x)$) are estimated by Maximum Likelihood (ML). To find the ML estimates for the model parameters, Latent Gold uses both the Expectation-Maximization and the Newton-Raphson algorithm (Vermunt and Magidson 2013a).

In Equation (3), only the indicators are used to estimate the likelihood of being in a specific true category. However, it is also possible to make use of covariate information to estimate the LC model. The third assumption we then make is that the measurement errors are independent of the covariates. An example of a covariate which can help in identifying whether someone owns or rents a house is *marital status*, this covariate is denoted by Q and can be added to Equation (3):

$$P(\mathbf{Y} = \mathbf{y}|Q = q) = \sum_{x=1}^{C} P(X = x|Q = q) \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{4}$$

Covariate information can also be used to impose a restriction on the model, to make sure that the model does not create a combination of a category of the "true" variable and a score on a covariate that is in practice impossible. For example, when an LC model is estimated to measure the variable *home ownership* using three indicator variables and a covariate (denoted by Z) measuring *rent benefit*, the impossible combination of owning a house and receiving rent benefit should not be created.

Throughout the article, we compare four approaches that researchers might administer when performing analyses using composite data sets containing classification errors and edit restrictions. In the first approach, researchers completely ignore the composite data structure and directly use one variable (which measures a construct that is measured by other variables in the composite data set as well) to obtain estimates of interest, for example a cross-table proportion or a logistic regression coefficient. In the second approach, researchers use an LC model to correct for classification errors, but are not aware of the edit restriction. The LC model used in this approach is equal to Equation (4); we call this the *unconditional model*. In the third approach, researchers are aware of the edit restriction, but they assume that including the restriction covariate (Z) in the LC model is enough to account for this; they do not explicitly mention the restriction itself. We call this the *conditional model*:

$$P(\mathbf{Y} = \mathbf{y}|Q = q, Z = z) = \sum_{x=1}^{C} P(X = x|Q = q, Z = z) \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{5}$$

Only in the fourth approach, the restriction is imposed directly in the LC model to fix the cell proportion of the impossible combination to 0; we call this the *restricted conditional model*. In the example where Z measures *rent benefit*, and the latent "true" variable measures *home ownership*, the imposed restriction is:

$$P(X = \text{own}|Z = \text{rent benefit}) = 0. \tag{6}$$

By using such a restriction, we can take impossible combinations with other variables into account, while we estimate an LC model for the underlying true measure. The restriction is imposed by specifically denoting which cell in the cross-table between the covariate and the latent variable should contain zero observations and giving this cell a weight of 0, resulting in constrained estimation (Vermunt and Magidson 2013b).

By specifying a model as in Equation (4) or in Equation (5), we assume that the covariate measure is in fact error-free, which is the fourth assumption we make. A fifth

assumption is that the edit rules applied are hard edit rules,in contrast to soft edit rules where there is a small probability that the edit is in fact possible. These five assumptions (assumption that $P(\mathbf{Y} = \mathbf{y})$ is a weighted average of $P(\mathbf{Y} = \mathbf{y}|X = x)$; assumption of local independence; assumption that measurement errors are independent of covariates; assumption that the covariate is error-free; assumption of hard edits) are specific for the LC model we use.

However, in practice it is very likely that one of these assumptions is not met. For example, with the assumption of local independence, we assume that when a mistake is made in one indicator, this is unrelated to the answers on other indicators. This assumption is probably met when one indicator originates from a survey and another from a register. If two indicators both originate from surveys, it is much more likely that a respondent makes the same mistake in both surveys, this assumption would then not be met. We can also think of situations where the assumption that misclassification is independent of covariates is not met. For example with tax registration by businesses, the number of delays and mistakes tends to be related to company size, since appropriate administration is better institutionalized in larger companies. The assumption that a covariate is free of error is in practice almost never met, since all sources always contain some error. The last assumption made is that the edits applied are hard edits. In some cases soft edits might be more appropriate, for example when a combination of scores is highly unlikely but not impossible, such as the combination of being ten years old and having graduated from high school.

Luckily these assumptions can be relaxed by specifying more complex LC models. However, whether you are able to relax these assumptions depends on your specific data structure. More specifically, it depends on whether your model is still identifiable. Unfortunately, model identifiability is not straightforward. For example, a model with three dichotomous indicators is identifiable, while a model with two dichotomous indicators is not. Adding a covariate to this model would make it identifiable. Adding a restriction to a model can also help to make an unidentifiable model identifiable. Since it is not possible to present general recommendations here, we refer to Biemer (2011) for more information about model identifiability. Examples of complex latent variable models which incorporate the different assumptions discussed in official statistics data sets are Pavlopoulos and Vermunt (2015) and Scholtus and Bakker (2013). Model identification can be checked in Latent Gold by assessing whether the Jacobian of the likelihood is full rank at a larger number of random parameter values (Forcina 2008). All models in this article were confirmed to be identifiable.

How missing values in the indicators and covariates are handled is also dependent on model specification. We specified the model as such that the indicators are part of the estimation procedure. Missing values are therefore handled by Full Information Maximum Likelihood (FIML) (Vermunt and Magidson 2013b, 51–52). Covariates are treated as fixed and listwise deletion will be applied to missing values here.

By applying Bayes' rule to the LC models from Equation (4), Equation (5), or Equation (6), posterior membership probabilities can be obtained. These posterior membership probabilities represent the probability of being in an LC given a specific combination of scores on the indicators and covariates ($P(X = x|Y = y, Q = q, Z = z)$).

For example, the posterior membership probabilities for the *conditional model* are obtained by:

$$P(X = x|Y = y, Q = q, Z = z) = \frac{P(X = x|Q = q, Z = z)\prod_{l=1}^{L} P(Y_l = y_l|X = x)}{\sum_{x=1}^{C} P(X = x|Q = q, Z = z)\prod_{l=1}^{L} P(Y_l = y_l|X = x)}. \quad (7)$$

These posterior membership probabilities can be used to impute latent variable X. To distinguish between the unobserved latent variable X, described by the LC model, and the variable after imputation, we denote this imputed variable by W. Different methods exist to obtain W. An example is modal assignment, where each respondent is assigned to the class for which its posterior membership probability is the largest. To correctly incorporate uncertainty caused by the classification errors, we use multiple imputation to estimate W. We first create m empty variables $(W_1, \ldots, W_m)$ and we impute them by drawing one of the LCs by sampling from the posterior membership probabilities from the m LC models.

With the *restricted conditional model*, we want to make sure that cases are not assigned to categories on the latent "true" variable which would result in impossible combinations with scores on other variables, such as the combination "rent benefit" × "own". Therefore, the restriction set in Equation (6) is also used here.

After we created m variables by imputing them using the posterior membership probabilities obtained from each of the m LC models, the estimates of interest can be obtained. For example, we can be interested in a cross table between imputed "true" variable W and covariate Z, where our estimate of interest $\hat{\theta}$ can be the cell proportion $P(W = 1, Z = 1)$. The m estimates of $\hat{\theta}$ can now be pooled by making use of the rules defined by Rubin for pooling (Rubin 1987, 76). The pooled estimate is obtained by

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i. \quad (8)$$

The total variance is estimated as

$$\mathrm{VAR}_{\mathrm{total}} = \overline{\mathrm{VAR}}_{\mathrm{within}} + \mathrm{VAR}_{\mathrm{between}} + \frac{\mathrm{VAR}_{\mathrm{between}}}{m}, \quad (9)$$

where $\overline{\mathrm{VAR}}_{\mathrm{within}}$ is the within imputation variance calculated by

$$\overline{\mathrm{VAR}}_{\mathrm{within}} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{VAR}_{\mathrm{within}_i}. \quad (10)$$

$\mathrm{VAR}_{\mathrm{within}_i}$ is estimated as the variance of the proportion of $\hat{\theta}_i$,

$$\frac{\hat{\theta}_i \times (1 - \hat{\theta}_i)}{N}, \quad (11)$$

where N is the number of units in the composite data set, and $\mathrm{VAR}_{\mathrm{between}}$ is calculated by

$$\mathrm{VAR}_{\mathrm{between}} = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta})(\hat{\theta}_i - \hat{\theta})'. \quad (12)$$

Besides the uncertainty caused by missing or conflicting data represented by the spread of parameter estimate values, $VAR_{between}$ also contains parameter uncertainty, which was introduced by the bootstrap performed in the first step of the MILC method.

3. Simulation

3.1. Simulation Approach

To empirically evaluate the performance of MILC, we conducted a simulation study using R (R Core Team 2014). We start by creating a theoretical population using Latent Gold (Vermunt and Magidson 2013a) containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) measuring the latent dichotomous variable (X); one dichotomous covariate (Z) which has an impossible combination with a score of the latent variable; and one other dichotomous covariate (Q). The theoretical population is generated using the restricted conditional model. When samples are drawn, it can happen that the LC model estimated from a sample assigns a non-zero probability to an impossible combination, so these errors are due to sampling. Furthermore, variations are made in the generated data sets according to scenarios described in the following sections.

When evaluating an imputation method, the relation between the imputed latent variable and other variables should be preserved since these relations might be the subject of research later on. When investigating the performance of MILC, there are two relations we are particularly interested in. We are interested in the relation between the imputed latent variable W and the covariate Z, which has an impossible combination with a score on the latent variable. The four cell proportions of the 2×2 table are denoted by: $W1 \times Z_1$, $W_2 \times Z_1$, $W_1 \times Z_2$ and $W_2 \times Z_2$. The cell $W_1 \times Z_2$ is the impossible combination, and should contain 0 observations. We compare the cell proportions of a 2×2 table of the population latent variable X and Z with the cell proportions of a table of the imputed latent variable W and Z from the samples. Furthermore, we are interested in the relation between W and covariate Q. To investigate this relation, we compare the coefficient of a logistic regression of the latent population variable X on Q with the logistic regression coefficient of the imputed W regressed on Q.

To investigate these relations, we look at three performance measures. First, we look at the bias of the estimates of interest. The bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the estimate over all replications. To confirm that the standard errors of the estimates were properly estimated, the ratio of the average standard error of the estimate over the standard deviation of the 1,000 estimates was also examined.

We expect the performance of MILC to be influenced by the measurement quality of the indicators, the marginal distribution of covariates Z and Q, the sample size, and the number of multiple imputations. The quality of the indicators is represented by classification probabilities. They represent the probability of a specific score on the indicator given the latent class. If the quality of the indicators is low, it will be more difficult for MILC to assign cases to the correct latent classes.

From Geerdinck et al. (2014) we know that classification probabilities of 0.95 and higher can be considered realistic for population registers. Pavlopoulos and Vermunt (2015) detected a classification probability of 0.83 in the Dutch Labour Force Survey. We investigate a range of classification probabilities around the values found, from 0.70 to 0.99. The marginal distribution of Z, $P(Z)$, is also expected to influence the performance of MILC. A higher value for $P(Z = 2)$ can give, for example, more information to the latent class model to assign scores to the correct latent class. Sample size may influence the standard errors and thereby the confidence intervals. The performance of MILC can also depend on the number of multiple imputations. Investigation of several multiple imputation methods have shown that five imputations are often sufficient (Rubin 1987). However, with complex data, it can be the case that more imputations are needed. As a result, the simulation conditions can be summarized as follows:

- Classification probabilities: 0.70; 0.80; 0.90; 0.95; 0.99.
- $P(Z = 2)$: 0.01; 0.05; 0.10; 0.20.
- Sample size: 1,000; 10,000.
- Logit coefficients of X regressed on Q of $\log(0.45/(1 - 0.45) = -0.2007$, $\log(0.55/(1 - 0.55) = 0.2007$ and $\log(0.65/(1 - 0.65) = 0.6190$ corresponding to estimated odds ratio of 0.81, 1.22 and 1.86. The intercept was fixed to 0
- Number of imputations: 5; 10; 20; 40.

To illustrate the measurement quality corresponding to different conditions, Figure 2 shows the entropy R^2 of the models under different values for $P(Z = 2)$ and classification probabilities. The entropy indicates how well one can predict class membership based on

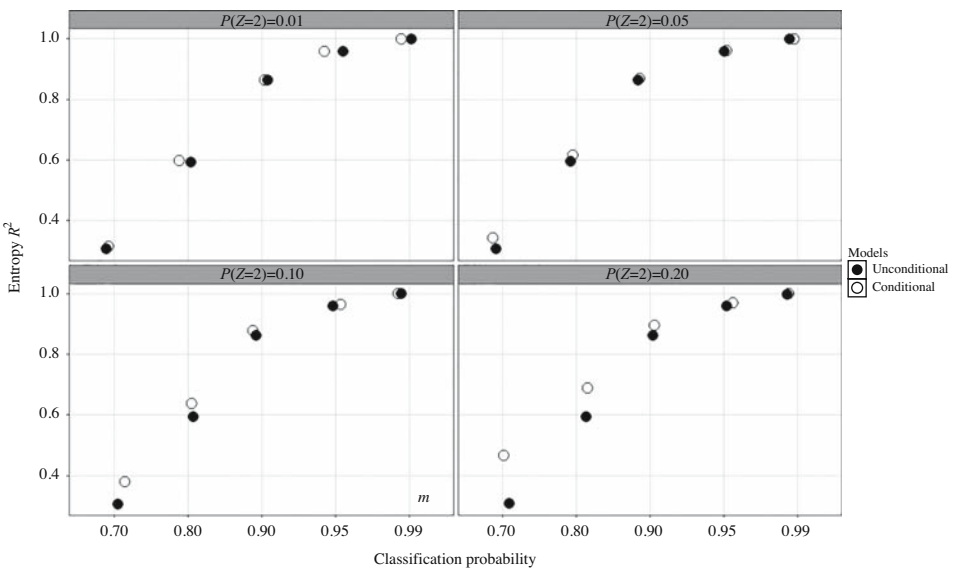

Fig. 2. Entropy R^2 of the unconditional and conditional model with different values for the classification probability and $P(Z = 2)$. The restricted conditional model has the same entropy R^2 as the conditional model because the models contain the same variables.

the observed variables, and is measured by:

$$EN(\alpha) = -\sum_{j=1}^{N}\sum_{x=1}^{X}\alpha_{jx}log\alpha_{jx}, \tag{13}$$

where α_{jx} is the probability that observation j is a member of class x, and N is the number of units in the composite data set. Rescaled with values between 0 and 1, entropy R^2 is measured by

$$R^2 = 1 - \frac{EN(\alpha)}{N\log X}, \tag{14}$$

where 1 means perfect prediction (Dias and Vermunt 2008). The *conditional* and the *restricted conditional model* have the same entropy R^2 because these models contain the same variables. All models with classification probabilities of 0.90 and above have a high entropy R^2 and are able to predict class membership well. When the classification probabilities are 0.70, the entropy R^2 is especially low. However, for the conditional and the restricted conditional model, the entropy R^2 under classification probability 0.70 increases as $P(Z = 2)$ increases. A larger $P(Z = 2)$ means that covariate Z contains more information for predicting class membership. Because covariate Z is not in the *unconditional model*, it makes sense that entropy R^2 remains stable for different values of $P(Z = 2)$ under this model. Furthermore, Figure 2 demonstrates that the performance of MILC is evaluated over an extreme range of entropy R^2 values and gives an indication of what we can expect from the MILC method under different simulation conditions.

3.2. Simulation Results

In this section we discuss our simulation results in terms of bias, coverage of the 95% confidence interval, and the ratio of the average standard error of the estimate over the standard deviation of the estimates. We do this in three sections. In the first section we discuss the 2 × 2 table of the imputed latent variable W and restriction covariate Z. In the second section, we investigate the relation between the imputed latent variable W and covariate Q. In the third section we investigate the influence of m, the number of bootstrap samples and multiple imputations. In the simulation results discussed in the first two sections, we used $m = 5$. When investigating the different simulation conditions, we focus on the performance of the four approaches discussed, using one indicator (Y_1), the *unconditional model*, the *conditional model* and the *restricted conditional model*. Interesting findings are illustrated with graphs containing results from situations when Y_1 is used and W is estimated using the restricted conditional model. For conditions that yielded approximately identical results, only one condition is shown in the figures. In Appendix A, tables with all results from the four approaches are given.

3.2.1. The Relation of Imputed Latent Variable W with Restriction Covariate Z

When we investigate the results in terms of bias (Figure 3), the restricted conditional model produces bias when the classification probabilities of the indicators are below 0.80. The bias of the cells where $P(Z = 1)$ for the restricted conditional model decreases when the classification probabilities increase or when $P(Z = 2)$ increases. This trend coincides

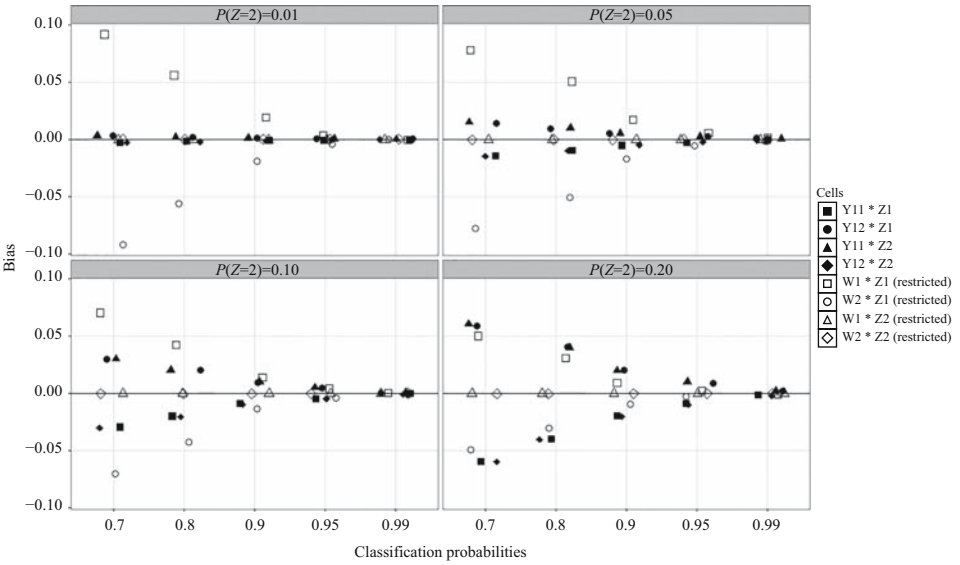

Fig. 3. Bias of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$. Sample size is 1,000 and $m = 5$.

with the trend we saw in Figure 2 for the entropy R^2, where a high entropy R^2 corresponds to a low bias. In contrast, when Y_1 is used, the bias of all cells is low when $P(Z = 2)$ is small, and increases as $P(Z = 2)$ increases. Furthermore, the restricted conditional model is the only model in which the cell representing the impossible combination $(W_1 \times Z_2)$ indeed contains 0 observations. $(Y1_1 \times Z_2)$ is never exactly 0.

When investigating the results for coverage of the 95% confidence intervals around the cell proportions (Figure 4), we see that the results differ over the different sample sizes. This is caused by the fact that even though the bias is not influenced by the sample size, the standard errors and therefore the confidence intervals are. Confidence intervals of biased estimates are therefore less likely to contain the population value. Furthermore, if the classification probabilities are larger,individuals are more likely to end up in the correct latent class, which also results in less variance, resulting in smaller confidence intervals. Confidence intervals cannot pe properly estimated for the impossible combination $Y1_1 \times Z_2$, since the proportions are very close to 0. This can be seen in Figure 4. Since $W_1 \times Z_2$ is not estimated with the restricted conditional model, confidence intervals cannot be estimated and coverage is therefore not shown.

The ratio of the average standard error of the estimate over the standard deviation of the simulated estimates tells us whether the standard errors of the estimates are properly estimated. In general, the values for both the situation of one indicator and the restricted conditional model, found in Figure 5, are both very close to 1. Only the standard errors for $W_1 \times Z_2$ are too small when one indicator is used. With the restricted conditional model, these are not estimated.

Overall, the small 2 × 2 cross tables investigated here containing a restriction covariate can be estimated when the LC model of the composite data set has an entropy R^2 of 0.90, or, when the sample size is large, an entropy R^2 of 0.95.

3.2.2. Relationship Between the Imputed Latent Variable W and Covariate Q

In the simulation results discussed in Subsection 3.2.1, the relation between the imputed latent variable W and covariate Z containing an impossible combination was investigated. Within the restricted conditional model, there was also another covariate, Q. We investigate the relation between W and Q with three different strengths of relations: intercepts are 0 and logit coefficients of W regressed on Q are -0.2007; 0.2007; 0.6190. Because the intercept is 0 in all conditions, we focus on the coefficients of Q when investigating the simulation results.

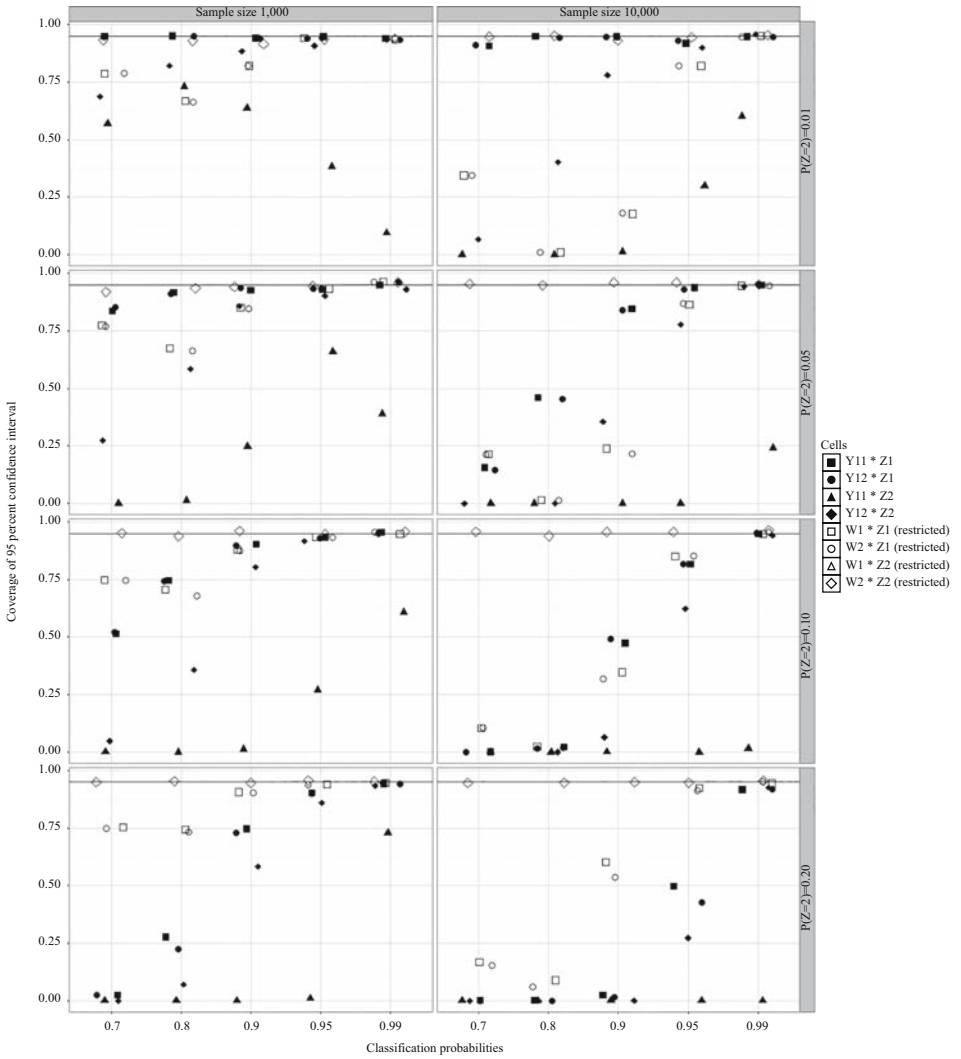

Fig. 4. Coverage of the 95% confidence interval of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$ and sample size, $m = 5$.

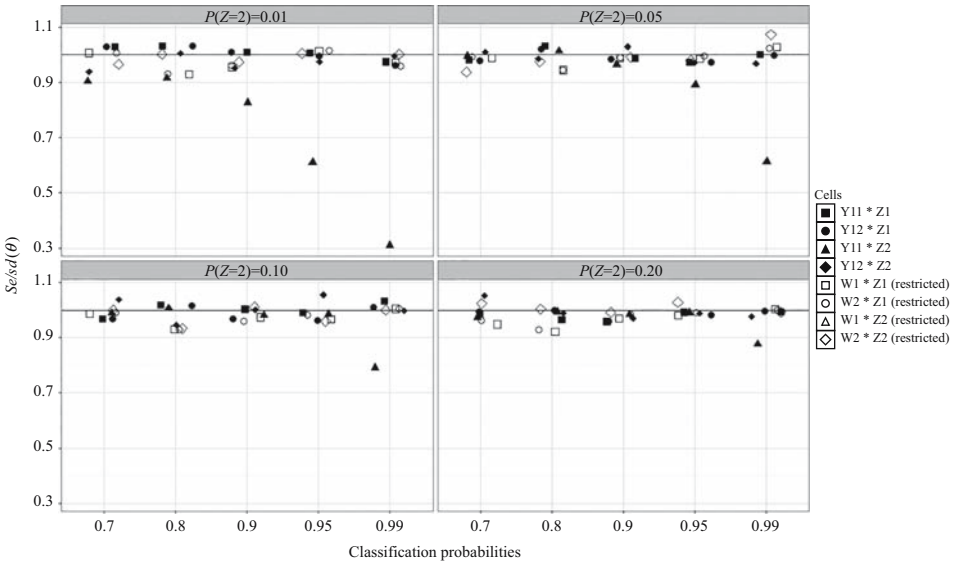

Fig. 5. $se/sd(\hat{\theta})$ of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$. Sample size is 1,000 and $m = 5$.

In Figure 6 we see that for the restricted conditional model, the bias is very close to 0 in all conditions. When Y_1 is used, the bias is much larger and is related to the classification probabilities.

In Figure 7 we see the results in terms of coverage of the 95% confidence interval. The conclusions we can draw here are comparable to the conclusions we drew from the results in terms of bias. When W is used (estimated using the restricted conditional model), the

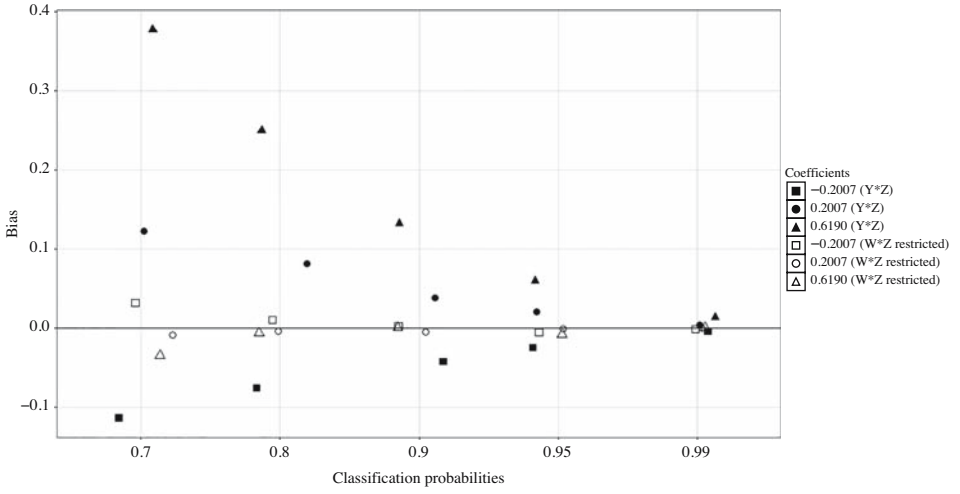

Fig. 6. Bias of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient and the classification probabilities. $P(Z = 2) = 0.01$, sample size is 1,000 and $m = 5$.

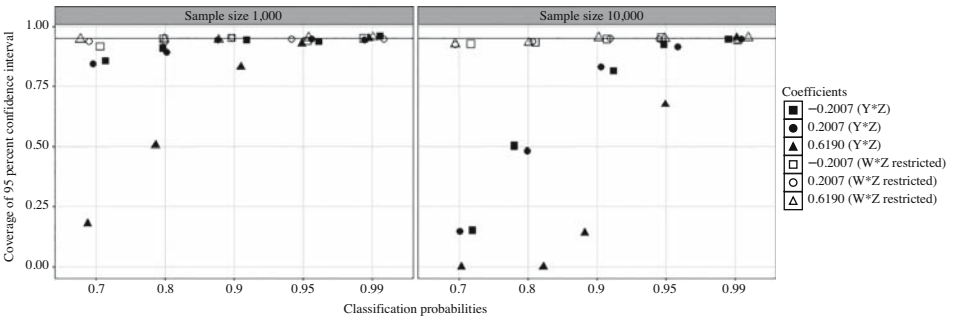

Fig. 7. Coverage of the 95% confidence interval of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient, the classification probabilities and sample size. $P(Z = 2) = 0.01$ and $m = 5$.

coverage of the 95% confidence is approximately 95 in all discussed conditions. When only one indicator (Y_1) is used, we see undercoverage when the population value of the logistic regression coefficient is 0.6190. This undercoverage is related to the classification probabilities and increases when the sample size increases. Results in terms of the ratio of the average standard error of the estimate over the standard deviation of the simulated estimates are very close to the desired ratio of 1. This is the case for all investigated simulation conditions, both when Y_1 is used or when W is used. Results are reported in Appendix A.

Overall, for the investigated conditions, unbiased estimates can be obtained when the LC model of the composite data set has an entropy R^2 of 0.60 or larger.

3.2.3. Number of Imputations

To investigate the effect of the number of bootstrap samples and imputations (m), we performed 5, 10, 20, and 40 bootstrap samples and imputations. The results of $m = 5$ and $m = 40$ can be found in Figure 8, while more results can be found in Appendix A. Both in terms of bias and coverage the MILC method performs equally well over the different numbers of m. It is important to note that the fraction of missing information corresponds, in the worst case, to the amount of missing data (Rubin 1987, 114). In our case, it depends on the entropy R^2, which is dependent on the classification and the covariates. Although

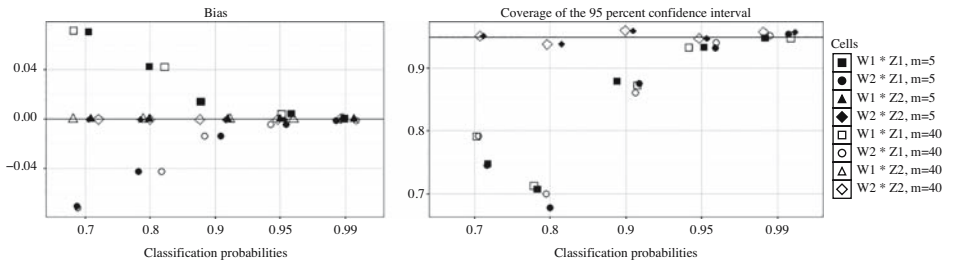

Fig. 8. Bias and coverage of the 95% confidence interval of four cells in the 2 × 2 table of covariate Z × W (estimated using the restricted conditional model). Number of bootstrap samples m = 5 and 40. The sample size is 1,000 and P(Z = 2) = 0.10.

the amount of missing values in W is 100%, the amount of missing information is much smaller when the entropy R^2 is larger than 0. This might explain why biased estimates with inappropriate coverage are obtained when the entropy R^2 is low, regardless of the size of m.

4. Application

4.1. Data

Home ownership is an interesting variable for social research. It has been related to a number of properties, such as inequality (Dewilde and Decker 2016), employment insecurity (Lersch and Dewilde 2015) and government redistribution (André and Dewilde 2016). Therefore, we apply the MILC method on a composite data set that brings together survey data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013 (Scherpenzeel 2011), which is administered by CentERdata (Tilburg University, The Netherlands) and a population register from Statistics Netherlands from 2013. Because samples for LISS were drawn by Statistics Netherlands, we were very well able to link these surveys and registers. From this composite data set, we use two variables indicating whether a person is either a home-owner or rents a house/other as indicators for the imputed "true" latent variable *home-owner/renter or other*. The composite data set also contains a variable measuring whether someone receives rent benefit from the government. A person can only receive rent benefit if this person rents a house. In a cross-table between the imputed latent variable *home-owner/renter* and *rent benefit*, there should be 0 persons in the cell "home-owner *s* receiving rent benefit". If people indeed receive rent benefit and own a house, this could be interesting for researchers and requires investigation. A more detailed LC model should then be specified, modelling local dependencies and allowing for error in the variable 'rent benefit'. However, this is outside the scope of the present study. We assume this to be measurement error, and therefore want this specific cell to contain 0 persons. Research has previously been done regarding the relation between home ownership and marital status (Mulder 2006). A research question here could be whether married individuals more often live in a house they own compared to non-married individuals. Therefore, a variable indicating whether a person is married or not is included in the latent class model as a covariate. The three data sets used to combine the data are discussed in more detail below:

- **Registration of addresses and buildings (BAG):** A register with data on addresses containing information about its buildings, owners and inhabitants originating from municipalities from 2013. Register information is obtained from persons who filled in the LISS studies and who declared that we are allowed to combine their survey information with registers. In total, this left us with 3,011 individuals. From the BAG we used a variable indicating whether a person "owns"/"rents"/"other" the house he or she lives in. Because our research questions mainly relate to home-owners, we recoded this variable into "owns"/"rents or other". This variable does not contain any missing values.
- **LISS background study:** A survey on general background variables from January 2013. From this survey we also have 3,011 individuals. We used the variable *marital*

status, indicating whether someone is "married"/"separated"/"divorced"/"widowed"/"never been married". As we are only interested in whether a person is married or not, we recoded this variable in such a way that "married" and "separated" individuals are in the recoded "married" category, and the "divorced", "widowed" and "never been married" individuals are in the "not married" category. It is difficult to handle a category as "separated" in such a situation. However, separated individuals are technically still married. Although they can in theory be more likely to live out of the registered address, it is difficult to make assumptions and therefore we decided to recode them into the category "married". This variable did not contain any missing values. We also used a variable indicating whether someone is a "tenant"/"sub-tenant"/"(co-) owner"/"other". We recoded this variable in such a way that we distinguish between "(co-) owner" and "(sub-) tenant or other". This variable had 14 missing values.

- **LISS housing study:** A survey on housing from June 2013. From this survey we used the variable *rent benefit*, indicating whether someone "receives rent benefit"/"the rent benefit is paid out to the lessor"/"does not receive rent benefit"/"prefers not to say". Because we are not interested in whether someone receives the rent benefit directly or indirectly, we recoded the first two categories into "receiving rent benefit". No one selected the option "prefers not to say". For this variable, we had 2,232 missing values resulting in 779 observations. The number of observations is small, because a selection variable (indicating whether someone rents their house) was used in the survey. Dependent interviewing has been used here. Only the individuals indicating that they rent their house in this variable were asked if they receive rent benefit. This selection variable could also have been used as an indicator in our LC model. However, because of the strong relation between this variable and the rent benefit variable we decided to leave it out of the model.

These data sets are linked at person level where matching is done on person identification numbers. In addition, matching could also have been done on date, since the surveys were conducted at different time points within 2013. However, mismatches on dates are a source of measurement error, and are therefore left in for illustration purposes. Although it is not necessarily the case in practice, the assumption is made that the covariate 'rent benefit' is measured without error, so we are able to apply the LC model investigated in the simulation study in practice. In Table 1, it can be seen that 48 individuals rent a home according to the BAG register, while stating to own a home in the LISS background survey. Furthermore, 155 individuals own a home according to the BAG register, while stating that they rent a home in the LISS background survey.

Table 1. Cross-table between the own/rent variable originating from the LISS background survey and the own/rent variable originating from the BAG register.

		Register	
		Rent	Own
Background survey	Rent	902	155
	Own	48	1,892

Not every individual is observed in every data set. This causes some missing values to be introduced when the different data sets are linked at a unit level. These records are not missing, but they are considered as non-sampled individuals. Full Information Maximum Likelihood was used to handle the missing values in the indicators (Vermunt and Magidson 2013b, 51–52).

The MILC method is applied to impute the latent variable *home owner/renter* by using two indicator variables and two covariates and the *restricted conditional model*. For results when the *unconditional* and the *conditional* model are applied we refer to Appendix B. In Table 2 classification statistics about the model is given, indicating how we can compare the results of this model to the information we obtained in the simulation study. Both the entropy R^2 and the classification probabilities are comparable to conditions we tested in the simulation study and in which the MILC method appeared to work very well. The classification probabilities for the LISS background survey and the BAG register indicate that theyboth have a high quality, but are error prone. Furthermore, P(married) and P(rent benefit) cannot be compared directly to the set up of the simulation study, but information provided by the covariates is taken into account in the entropy R^2.

For the two variables measuring home ownership, we can see from the cell totals in Table 3 whether individuals who say to own their home also receive rent benefit, which is not allowed. However, in practice these discrepancies can be caused by the fact that people make mistakes when filling in a survey, or for example because people were moving during the period the surveys took place. Furthermore, the total number of individuals who can be found in the table of the LISS background study are only 779, and for the BAG register 772. This is because only the people indicating that they rented a house in the LISS Housing study were asked the question whether they received rent benefit. For the LISS background study we see that eight individuals are in the cell representing the impossible combination of owning a house and receiving rent benefit, and for the BAG register 4. If we investigate the cell proportions estimated by the MILC method, we see that both the conditional and the unconditional model replicate the structure of the indicators very well, but that individuals are still assigned to the cell of the impossible combination (see Appendix B). To get this correctly estimated, we need the restricted conditional model. The marginals of the variable *own/rent* (in the upper block of Table 3) for the different models are all very close to each other, and closer to the estimates in the BAG register than to the estimates of the LISS background study. Also note that individuals with missing

Table 2. Entropy R^2 of the restricted conditional model; classification probabilities of the indicators and marginal probabilities of the covariates. The covariate rent benefit takes information of 779 individuals into account and marital status variable of 3,011 individuals.

			Restricted conditional model
Entropy R^2			0.9380
Classification probability	LISS background	P(rent|LC rent)	0.9344
		P(own|LC own)	0.9992
	BAG register	P(rent|LC rent)	0.9496
		P(own|LC own)	0.9525
P(rent benefit)			0.3004
P(married)			0.5284

Table 3. The first block represents the (pooled) marginal proportions of the variable own/rent. The second block represents the (pooled) proportions of the variable own/rent for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable own/rent for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last row represents the restricted conditional model used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
LISS background	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
Restricted conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
	P(own × rent benefit)		P(rent × rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
Restricted conditional	0.0000	-	0.2978	[0.2649; 0.3307]
	P(own × no rent benefit)		P(rent × no rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
Restricted conditional	0.0213	[−0.0116; 0.0542]	0.6773	[0.6444; 0.7102]

values on the variable *rent benefit* are not taken into account in the 2 × 2 table of *rent benefit* × *own/rent*.

After we investigated the cross table between home ownership and rent benefit, we were also interested in whether marriage can predict home ownership. When we consider the BAG register, we see that the estimated odds of owning a home when not married are $e^{-1.2331} = 0.29$ times the odds when married, while they are $e^{-1.3041} = 0.27$ when the LISS background survey is used. It is interesting to see that when the restricted conditional MILC model is used to obtain an estimate that also corrects for the impossible combination of owning a house and receive rent benefit, we see that this coefficient is even a little less strong, namely $e^{-1.3817} = 0.25$. Overall, these results show us that although non-married individuals are approximately equally likely to own or rent a house, married individuals are three times more likely to own a house than to rent one.

5. Discussion

In this article we introduced the MILC method, which combines latent class analysis with edit restrictions and multiple imputation to obtain estimates for variables of which we had multiple indicators in a composite data set. We distinguished between invisibly present and visibly present errors (commonly solved by edit restrictions), and argued the need for a method that takes them into account simultaneously. We evaluated the MILC method in terms of its ability to correctly take impossible combinations and relations with other

Table 4. The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The third row represents the restricted conditional model used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the variable owning/renting a house.

	Intercept		Marriage	
	Estimate	95% CI	Estimate	95% CI
BAG register	2.4661	[2.2090; 2.7233]	−1.2331	[−1.3901; −1.0760]
LISS background	2.7620	[2.4896; 3.0343]	−1.3041	[−1.4678; −1.1405]
Restricted conditional	2.7712	[2.5036; 3.0389]	−1.3817	[−1.6493; −1.1140]

variables into account. We assessed these relations by investigating the bias of $\hat{\theta}$, coverage of the 95% confidence interval, and $se/sd(\hat{\theta})$ in different conditions in a simulation study. The performance of MILC appeared to be mainly dependent on the entropy R^2 value of the LC model. We conclude that a different quality of the composite data set is required to obtain unbiased estimates and standard errors for different types of estimates. In cases of 2×2 tables including an edit restriction, a higher quality of the composite data set was required (entropy R^2 of 0.90), while unbiased estimates and standard errors for logit coefficients can already be obtained with an entropy R^2 value of 0.60.

An example of a composite data set containing data from the LISS panel and the BAG register were shown to have adequate entropy R^2 and we investigated the MILC method using the unconditional model, the conditional model and the restricted conditional model. All models can potentially be used when using the MILC method in practice. However, if there are edit restrictions within the data that need to be taken into account, only the restricted conditional model is appropriate. In light of our main findings, the MILC method can be seen as an alternative for methods previously used for handling visibly and invisibly present errors. This was done either separately using latent variable models and edit rules, or simultaneously by Manrique-Vallier and Reiter (2016), by using one file and one measurement.

A number of limitations of the current study are related to the assumptions we made when specifying the LC model. We assumed that the observed indicators were independent of each other given a unit's score on the latent variable, which means that when a mistake is made on an indicator originating from one source, this is independent of mistakes made on indicators from other sources. For example, if multiple indicators originate from comparable surveys, there is a probability that a respondent makes the same mistake in both surveys; this assumption is then not met. There are ways to relax this assumption by extending the LC model, but we did not investigate the performance of the MILC method if this assumption is relaxed. We also assumed that the misclassification is independent of the covariates. This is also an assumption that in some cases should be relaxed, which we did not investigate as well. Furthermore, the assumption was made that the covariates are free of error. Since this assumption is often not met, ways to relax this assumption should be investigated as well as the performance of the MILC method in such cases. Finally, it was assumed that all edits applied were hard edits, while sometimes soft edits are better applicable. We applied the edits by specifying which cell in the cross table between the latent variable and a covariate should have a weight of 0, while it is also

possible to fix the relevant logit parameter to a very small number. In this way, it should be possible to apply hard or soft edit restrictions. However, we did not investigate the performance of the MILC method when edits are specified in such a manner. We also did not investigate the performance of the LC model used here when some of the previously discussed assumptions are not met.

If a researcher is interested in investigating the relationship between the imputed latent variable and many other variables, all these variables should be included in the LC model as covariates. With the LC three-step approach (Bakk et al. 2016), relationships between the imputed latent variable and other variables (not incorporated in the LC model) can be investigated as well. Edit restrictions could then be added later on as well. However, this three-step approach has not been incorporated in the MILC framework. More investigation can also be done on how the MILC framework handles missing values within covariates, linkage errors and selection errors. Furthermore, the current simulation study only considers dichotomous variables. The current simulation study shows how the method works and it gives some indications of when the method works. This simulation was also comprehensive enough to discover the relation between the quality of the results after imputation and the entropy R^2 value of the LC model. However, it should still be investigated if this relationship holds with larger numbers of indicators, covariates and larger numbers of edit restrictions, and what the exact limitations will be. Also situations when indicators have different numbers of categories are not yet investigated.

Another point of discussion is that we used three indicators in our LC model. In practice, it is more likely that researchers find only two indicators for an underlying true measure in their composite data set. However, a model with two indicators is not identifiable so an additional covariate is necessary. The fact that we used three indicators might seem like a disadvantage. However, a three indicator model and a two indicator plus covariate model are Markov equivalent, which means that they yield the same set of conditional inference assumptions and an identical likelihood.

It should also be noted that MILC can be applied to indicators coming from both population registers and sample surveys. When the indicators only come from sample surveys, we can use the standard rules for pooling as defined by Rubin (1987). However, when at least one of the indicators is sourced from a complete population register, we can choose to either only impute the survey variables, and weigh them to appropriately represent the population variables, or we can choose to impute both the survey and population variables, and use adjusted rules for pooling (Vink and van Buuren 2014). We use these adjusted rules because in the case of register indicators all sampling variability is captured by the between imputation variance, so the within variance should be left out of the equation. In this article, we consider the situation where samples and population registers are linked at a unit level, resulting in a composite data set consisting of only the individuals that were also in the survey sample. However, it is important to be aware of necessary adjustments when population registers are used.

Appendix A

- **Table 1** *Y*1 × *Z*: This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of *Y*1 and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N*), number of bootstrap samples, $m = 5$.
- **Table 2** *W* × *Z* **unconditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the unconditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N*), number of bootstrap samples, $m = 5$.
- **Table 3** *W* × *Z* **conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the conditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N*), number of bootstrap samples, $m = 5$.
- **Table 4** *W* × *Z* **restricted conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the restricted conditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N*), number of bootstrap samples, $m =$.
- **Table 5** *Y*1 × *Q*: This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of Y_1 on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.
- **Table 6** W × Q **unconditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the unconditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.
- **Table 7** *W* × *Q* **conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the conditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.
- **Table 8** W × Q **restricted conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the restricted conditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.

- **Table 9** W × Z **restricted conditional** *m***:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W (imputed using the restricted conditional model) and covariate Z with classification probabilities 0.90, $P(Z = 2) = 0.1$, sample size = 1000, 00 and different values for m.

Table 1. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of Y_1 and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and different values for sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	−.0030	.9490	1.0282	−.0020	.9490	1.0315	−.0011	.9400	1.0072	−.0009	.9460	1.0053	−.0006	.9370	0.9730
		2	.0031	.9480	1.0301	.0020	.9480	1.0312	.0011	.9390	1.0089	.0009	.9410	0.9946	.0005	.9350	0.9611
		3	.0030	.5690	0.9032	.0019	.7290	0.9144	.0010	.6370	0.8257	.0005	.3820	0.6093	.0001	.0950	0.3097
		4	−.0031	.6880	0.9381	−.0020	.8210	1.0064	−.0010	.8840	0.9526	−.0005	.9070	0.9748	−.0000	.9330	0.9967
	.05	1	−.0146	.8340	0.9803	−.0096	.9150	1.0300	−.0055	.9250	0.9860	−.0030	.9280	0.9716	−.0008	.9470	0.9994
		2	.0144	.8530	0.9799	.0097	.9120	1.0207	.0052	.9370	0.9843	.0027	.9330	0.9723	.0010	.9630	0.9980
		3	.0150	.0000	0.9944	.0099	.0120	1.0129	.0050	.2480	0.9638	.0025	.6600	0.8902	.0005	.3880	0.6134
		4	−.0147	.2740	1.0104	−.0100	.5850	0.9858	−.0047	.8560	1.0292	−.0022	.9020	0.9713	−.0007	.9300	0.9680
	.10	1	−.0297	.5130	0.9659	−.0202	.7450	1.0181	−.0093	.9010	1.0017	−.0051	.9320	0.9894	−.0004	.9540	1.0310
		2	.0297	.5210	0.9686	.0204	.7410	1.0152	.0093	.8970	0.9680	.0047	.9300	0.9626	.0002	.9520	1.0109
		3	.0300	.0000	0.9885	.0202	.0000	1.0051	.0099	.0110	0.9797	.0050	.2690	0.9833	.0010	.6070	0.7905
		4	−.0300	.0480	1.0377	−.0204	.3570	0.9469	−.0099	.8030	1.0010	−.0046	.9150	1.0553	−.0008	.9460	0.9981
	.20	1	−.0600	.0240	0.9841	−.0400	.2760	0.9638	−.0198	.7460	0.9571	−.0092	.9020	0.9908	−.0018	.9420	0.9911
		2	.0593	.0270	0.9931	.0405	.2260	0.9983	.0203	.7310	0.9611	.0090	.8990	0.9822	.0020	.9430	0.9949
		3	.0605	.0000	0.9737	.0399	.0000	0.9931	.0199	.0000	0.9830	.0100	.0100	0.9891	.0020	.7300	0.8763
		4	−.0597	.0000	1.0523	−.0404	.0710	0.9887	−.0204	.5840	0.9690	−.0098	.8610	0.9887	−.0022	.9350	0.9769

Table 1. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
10,000	.01	1	−.0031	.9060	1.0085	−.0022	.9470	1.0526	−.0009	.9460	0.9997	−.0006	.9170	0.9558	−.0000	.9460	0.9592
		2	.0031	.9090	1.0003	.0021	.9440	1.0559	.0009	.9450	1.0010	.0006	.9310	0.9685	.0000	.9470	0.9599
		3	.0030	.0000	1.0166	.0020	.0000	0.9763	.0010	.0130	0.9765	.0005	.2980	0.9257	.0001	.6010	0.7449
		4	−.0030	.0670	1.0061	−.0020	.4020	1.0053	−.0010	.7800	0.9837	−.0005	.8990	1.0001	−.0001	.9570	1.0448
	.05	1	−.0149	.1550	0.9830	−.0102	.4580	0.9722	−.0049	.8440	1.0069	−.0024	.9360	1.0488	−.0005	.9460	0.9902
		2	.0148	.1460	0.9879	.0103	.4550	0.9905	.0048	.8380	0.9870	.0024	.9290	1.0282	.0005	.9530	1.0145
		3	.0150	.0000	0.9912	.0100	.0000	0.9905	.0050	.0000	0.9780	.0025	.0000	1.0130	.0005	.2400	1.0204
		4	−.0150	.0000	1.0053	−.0101	.0000	0.9859	−.0049	.3550	1.0221	−.0025	.7780	0.9873	−.0005	.9410	0.9776
	.10	1	−.0298	.0000	1.0235	−.0200	.0210	0.9755	−.0101	.4720	1.0048	−.0052	.8140	0.9984	−.0010	.9460	1.0379
		2	.0298	.0000	1.0135	.0202	.0160	0.9866	.0100	.4910	1.0333	.0050	.8180	0.9866	.0010	.9510	1.0240
		3	.0300	.0000	1.0127	.0199	.0000	1.0195	.0100	.0000	0.9927	.0050	.0000	0.9951	.0010	.0140	1.0053
		4	−.0300	.0000	1.0421	−.0201	.0000	1.0141	−.0100	.0640	1.0627	−.0048	.6220	1.0058	−.0010	.9400	1.0022
	.20	1	−.0600	.0000	1.0166	−.0401	.0000	1.0019	−.0202	.0230	1.0247	−.0098	.4960	1.0247	−.0023	.9160	0.9981
		2	.0599	.0000	1.0085	.0400	.0000	1.0031	.0201	.0140	0.9854	.0099	.4280	1.0534	.0021	.9180	0.9838
		3	.0600	.0000	1.0068	.0401	.0000	0.9837	.0199	.0000	1.0067	.0100	.0000	1.0267	.0020	.0000	0.9471
		4	−.0599	.0000	0.9514	−.0399	.0000	1.0110	−.0199	.0000	0.9832	−.0100	.2730	1.0208	−0.0018	.9260	0.9964

Table 2. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the unconditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	.0366	.8470	1.0021	.0394	.7770	0.9261	.0173	.8320	0.9552	.0037	.9420	1.0147	.0006	.9480	1.0053
		2	−.0368	.8520	1.0019	−.0393	.7780	0.9289	−.0173	.8370	0.9612	−.0037	.9430	1.0136	−.0005	.9510	1.0140
		3	.0027	.9370	1.2328	.0010	.8400	1.2466	.0002	.3170	0.8594	.0000	.0810	0.4324	.0000	.0040	0.0974
		4	−.0025	.8160	1.1160	−.0010	.9080	1.0406	−.0002	.9130	0.9804	−.0000	.9340	1.0045	−.0000	.9240	0.9772
	.05	1	.0249	.8540	1.0331	.0333	.8040	0.9736	.0153	.8770	0.9936	.0054	.9350	0.9894	.0016	.9620	1.0258
		2	−.0248	.8540	1.0347	−.0332	.8070	0.9743	−.0152	.8630	0.9943	−.0055	.9360	0.9994	−.0012	.9600	1.0235
		3	.0131	.6080	1.1421	.0049	.8320	1.3231	.0009	.7990	1.1260	.0002	.3140	0.7553	.0000	.0160	0.1506
		4	−.0132	.6620	1.0697	−.0050	.8740	1.0218	−.0010	.9320	0.9923	−.0001	.9430	0.9812	−.0005	.9590	1.0732
	.10	1	.0102	.8460	0.9645	.0253	.8430	0.9270	.0124	.8870	0.9782	.0041	.9340	0.9633	.0006	.9460	1.0048
		2	−.0102	.8450	0.9640	−.0250	.8380	0.9310	−.0120	.8820	0.9706	−.0039	.9330	0.9830	−.0009	.9520	1.0054
		3	.0260	.4920	1.0207	.0098	.5010	1.1634	.0017	.9490	1.3165	.0004	.4960	0.9755	.0000	.0270	0.2549
		4	−.0260	.5490	1.0473	−.0100	.7800	0.9656	−.0021	.9390	1.0089	−.0005	.9350	0.9589	.0003	.9580	1.0001
	.20	1	−.0215	.8540	0.9901	.0112	.8720	0.8983	.0076	.9240	0.9714	.0022	.9440	0.9852	−.0007	.9450	1.0061
		2	.0220	.8540	0.9917	−.0107	.8700	0.9106	−.0075	.9300	0.9971	−.0022	.9430	0.9964	.0007	.9420	1.0025
		3	.0501	.4130	1.0317	.0189	.2250	1.0623	.0031	.9330	1.3470	.0006	.7070	1.1819	.0000	.0620	0.3982
		4	−.0506	.4990	1.0398	−.0194	.7040	0.9939	−.0032	.9290	0.9898	−.0005	.9590	1.0322	.0000	.9530	0.9921

Table 2. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
10,000	.01	1	.0439	.7410	0.9406	.0420	.1180	0.9371	.0163	.2460	0.9242	.0049	.8280	0.9624	.0005	.9490	1.0112
		2	−.0440	.7390	0.9405	−.0420	.1160	0.9362	−.0163	.2560	0.9124	−.0049	.8260	0.9596	−.0005	.9480	1.0085
		3	.0025	.0170	1.2140	.0010	.2510	1.3434	.0002	.9530	1.2920	.0000	.4960	1.0528	.0000	.0200	0.2565
		4	−.0024	.3630	1.0981	−.0009	.8280	1.0447	−.0002	.9300	1.0030	−.0000	.9440	0.9849	−.0000	.9530	1.0248
	.05	1	.0314	.7980	0.9143	.0347	.1910	0.9367	.0143	.3370	0.9716	.0043	.8760	1.0095	.0001	.9450	0.9906
		2	−.0314	.7990	0.9117	−.0346	.1910	0.9476	−.0144	.3100	0.9818	−.0042	.8740	0.9970	−.0001	.9460	0.9851
		3	.0122	.0000	0.9927	.0047	.0000	1.2805	.0008	.3410	1.3722	.0001	.9330	1.3488	.0000	.1130	0.6631
		4	−.0123	.0170	1.0485	−.0047	.4330	1.0148	−.0007	.9470	1.0245	−.0002	.9540	1.0424	.0001	.9520	0.9968
	.10	1	.0172	.8690	0.9719	.0269	.3530	0.9559	.0123	.4490	0.9795	.0042	.8530	0.9596	.0001	.9440	0.9698
		2	−.0172	.8660	0.9749	−.0267	.3540	0.9753	−.0123	.4190	0.9746	−.0042	.8640	0.9899	−.0002	.9560	1.0132
		3	.0243	.0000	1.0485	.0092	.0000	1.2298	.0016	.0250	1.3530	.0003	.9630	1.4935	.0000	.1750	0.8093
		4	−.0243	.0030	1.0305	−.0094	.1400	1.0139	−.0016	.9100	1.0152	−.0003	.9580	1.0118	.0001	.9630	1.0389
	.20	1	−.0135	.8600	0.9420	.0111	.7430	0.9338	.0078	.7000	0.9701	.0025	.9220	0.9899	.0004	.9470	0.9896
		2	.0134	.8600	0.9490	−.0113	.7220	0.9505	−.0078	.6810	0.9894	−.0025	.9090	0.9983	−.0003	.9500	0.9825
		3	.0474	.0000	0.9731	.0177	.0000	1.0894	.0030	.0000	1.3102	.0005	.7840	1.4592	.0000	.2640	1.0452
		4	−.0473	.0010	0.9562	−.0175	.0180	0.9876	−.0030	.8760	0.9922	−.0005	.9470	0.9829	−.0001	.9550	1.0227

Table 3. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the conditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	−.2642	.4510	0.9192	.0544	.6940	0.9307	.0188	.8180	0.9471	.0037	.9420	0.9971	.0102	.9050	1.0478
		2	.2640	.4530	0.9210	−.0544	.6980	0.9304	−.0188	.8150	0.9519	−.0037	.9430	0.9962	−.0102	.9170	1.0569
		3	−.0025	.8190	1.1097	.0005	.5420	0.9409	.0001	.2030	0.7171	.0000	.0610	0.4060	−.0100	.0000	0.0746
		4	.0027	.9640	1.3693	−.0005	.9190	1.0043	−.0001	.9140	0.9741	−.0000	.9350	1.0036	.0100	.0100	1.0072
	.05	1	.1163	.6880	1.0093	.0543	.6570	0.9572	.0173	.8460	0.9963	.0055	.9310	0.9908	.0016	.9640	1.0268
		2	−.1162	.6890	1.0128	−.0543	.6400	0.9564	−.0171	.8420	0.9956	−.0056	.9330	0.9989	−.0011	.9620	1.0251
		3	.0055	.9170	1.0301	.0011	.7930	0.9647	.0002	.3500	0.8031	.0000	.0960	0.5438	.0000	.0150	0.1936
		4	−.0056	.8770	0.9839	−.0012	.9370	0.9854	−.0003	.9410	0.9926	.0000	.9430	0.9793	−.0005	.9590	1.0746
	.10	1	.1040	.6550	1.0122	.0474	.6600	0.9330	.0144	.8740	0.9752	.0042	.9320	0.9683	.0006	.9480	1.0060
		2	−.1040	.6480	1.0132	−.0472	.6510	0.9297	−.0140	.8610	0.9594	−.0041	.9360	0.9862	−.0009	.9560	1.0074
		3	.0084	.9190	0.9661	.0018	.8420	0.9172	.0002	.3530	0.7467	.0000	.0650	0.4029	.0000	.0160	0.2591
		4	−.0084	.8970	1.0497	−.0021	.9250	0.9437	−.0006	.9590	1.0080	−.0002	.9470	0.9595	.0002	.9570	1.0004
	.20	1	.0723	.6580	0.9574	.0354	.7010	0.9143	.0100	.9090	0.9647	.0024	.9450	0.9822	−.0009	.9440	1.0038
		2	−.0718	.6520	0.9722	−.0349	.6840	0.9238	−.0099	.9050	0.9785	−.0024	.9420	0.9975	.0008	.9420	1.0013
		3	.0118	.9290	0.9710	.0026	.9150	0.9981	.0003	.3880	0.7857	.0000	.0390	0.2919	.0000	.0250	0.3337
		4	−.0123	.9050	1.0121	−.0031	.9480	1.0107	−.0004	.9460	0.9904	.0000	.9580	1.0282	−.0000	.9550	0.9929

Table 3. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
10,000	.01	1	.1083	.3180	0.9304	.0580	.0140	0.9355	.0177	.1860	0.9253	.0050	.8180	0.9618	.0005	.9480	1.0122
		2	−.1084	.3150	0.9299	−.0581	.0140	0.9341	−.0178	.1750	0.9133	−.0050	.8140	0.9592	−.0005	.9480	1.0094
		3	.0006	.9300	0.9719	.0002	.8490	0.9536	.0000	.3510	0.7248	.0000	.1330	0.5662	.0000	.0180	0.2357
		4	−.0006	.9140	1.0145	−.0001	.9510	1.0016	−.0000	.9330	0.9943	−.0000	.9440	0.9827	−.0000	.9530	1.0248
	.05	1	.1036	.1730	0.9255	.0524	.0130	0.9195	.0159	.2330	0.9716	.0044	.8530	1.0045	.0001	.9450	0.9891
		2	−.1036	.1710	0.9222	−.0524	.0110	0.9285	−.0160	.2270	0.9780	−.0044	.8620	0.9924	−.0001	.9470	0.9840
		3	.0018	.8800	0.9726	.0004	.9500	0.9899	.0001	.5210	0.8260	.0000	.1170	0.4172	.0000	.0230	0.3223
		4	−.0018	.9020	1.0385	−.0005	.9430	0.9846	.0001	.9600	1.0233	−.0001	.9590	1.0413	.0001	.9520	0.9969
	.10	1	.0896	.0740	0.9529	.0455	.0180	0.9337	.0140	.3380	0.9850	.0043	.8500	0.9594	.0001	.9440	0.9693
		2	−.0897	.0730	0.9515	−.0453	.0170	0.9555	−.0140	.3200	0.9809	−.0043	.8570	0.9918	−.0002	.9560	1.0124
		3	.0026	.8500	0.9642	.0007	.9210	0.9950	.0001	.6890	0.8811	.0000	.1460	0.4519	.0000	.0080	0.1825
		4	−.0025	.9170	1.0370	−.0008	.9310	0.9986	−.0001	.9560	1.0123	−.0000	.9560	1.0116	.0001	.9640	1.0389
	.20	1	.0601	.0970	0.9126	.0312	.0760	0.9378	.0098	.5890	0.9694	.0027	.9240	0.9922	.0004	.9470	0.9894
		2	−.0602	.0910	0.9402	−.0313	.0570	0.9654	−.0097	.5310	0.9889	−.0027	.9080	1.0019	−.0003	.9510	0.9822
		3	.0036	.8420	0.9642	.0009	.9090	0.9216	.0002	.8180	1.0068	.0000	.2220	0.5693	.0000	.0020	0.0715
		4	−.0035	.8960	1.0040	−.0007	.9330	0.9912	−.0002	.9500	0.9984	−.0000	.9480	0.9811	−.0001	.9550	1.0229

Table 4. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the restricted conditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.01	1	.0920	.7870	1.0070	.0561	.6670	0.9301	.0190	.8210	0.9555	.0038	.9410	1.0149	−.0003	.9360	0.9726
		2	−.0922	.7880	1.0075	−.0561	.6630	0.9322	−.0190	.8220	0.9608	−.0038	.9420	1.0149	.0002	.9400	0.9595
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0002	.9330	0.9656	−.0001	.9290	1.0017	.0000	.9140	0.9738	.0000	.9360	1.0045	.0001	.9380	1.0022
	.05	1	.0780	.7740	0.9893	.0507	.6730	0.9450	.0171	.8520	0.9882	.0056	.9330	0.9864	.0016	.9630	1.0270
		2	−.0779	.7690	0.9927	−.0507	.6630	0.9456	−.0170	.8460	0.9896	−.0056	.9320	0.9962	−.0011	.9610	1.0249
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0001	.9200	0.9375	−.0001	.9360	0.9757	−.0001	.9410	0.9919	.0001	.9440	0.9800	−.0005	.9590	1.0741
	.10	1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663	.0006	.9480	1.0044
		2	−.0705	.7450	0.9894	−.0423	.6770	0.9283	−.0137	.8760	0.9607	−.0041	.9320	0.9822	−.0009	.9540	1.0052
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	.20	1	.0501	.7530	0.9482	.0309	.7440	0.9212	.0095	.9060	0.9686	.0024	.9410	0.9801	−.0007	.9470	1.0041
		2	−.0496	.7500	0.9611	−.0304	.7330	0.9279	−.0094	.9040	0.9845	−.0024	.9390	0.9907	.0007	.9430	1.0000
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0005	.9490	1.0238	−.0005	.9530	1.0044	−.0001	.9460	0.9915	.0001	.9560	1.0284	.0000	.9530	0.9923

Table 4. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
10,000	.01	1	.1021	.3450	0.9331	.0581	.0110	0.9398	.0178	.1780	0.9292	.0050	.8200	0.9664	.0005	.9500	1.0132
		2	−.1021	.3440	0.9330	−.0582	.0120	0.9391	−.0178	.1820	0.9168	−.0050	.8220	0.9633	−.0005	.9480	1.0101
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9470	0.9863	.0001	.9510	1.0032	.0000	.9310	0.9913	−.0000	.9440	0.9825	−.0000	.9530	1.0248
	.05	1	.0916	.2140	0.9175	.0513	.0140	0.9214	.0158	.2380	0.9697	.0044	.8630	1.0099	.0001	.9460	0.9894
		2	−.0916	.2120	0.9145	−.0512	.0130	0.9309	−.0159	.2170	0.9757	−.0044	.8690	0.9979	−.0001	.9470	0.9840
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0000	.9530	0.9966	−.0001	.9490	0.9824	.0001	.9600	1.0234	−.0001	.9590	1.0416	.0001	.9520	0.9971
	.10	1	.0794	.1050	0.9382	.0441	.0240	0.9314	.0139	.3470	0.9666	.0043	.8510	0.9603	.0001	.9460	0.9697
		2	−.0794	.1050	0.9361	−.0439	.0180	0.9537	−.0139	.3200	0.9653	−.0043	.8520	0.9925	−.0002	.9560	1.0129
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0000	.9580	1.0304	−.0002	.9390	0.9953	−.0000	.9560	1.0115	−.0000	.9560	1.0118	.0001	.9640	1.0389
	.20	1	.0531	.1680	0.9064	.0300	.0890	0.9311	.0096	.6020	0.9619	.0027	.9250	0.9907	.0004	.9450	0.9898
		2	−.0532	.1530	0.9333	−.0301	.0620	0.9565	−.0096	.5350	0.9772	−.0026	.9130	0.9982	−.0003	.9510	0.9831
		3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9460	0.9888	.0002	.9470	0.9933	−.0001	.9490	0.9993	−.0000	.9480	0.9803	−.0001	.9550	1.0229

Table 5. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of Y_1 regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	−.1139	.8550	1.0087	−.0763	.9080	1.0065	−.0432	.9430	0.9924	−.0253	.9360	0.9852	−.0047	.9590	1.0096
		.05	−.1254	.8360	1.0070	−.0756	.9220	1.0013	−.0503	.9230	0.9830	−.0170	.9290	0.9661	−.0025	.9500	0.9876
		.10	−.1156	.8620	1.0103	−.0823	.8970	0.9963	−.0335	.9550	1.0339	−.0228	.9420	0.9940	−.0066	.9460	0.9528
		.20	−.1144	.8610	1.0121	−.0762	.9030	0.9795	−.0378	.9430	1.0232	−.0157	.9500	1.0179	−.0063	.9480	0.9803
	.55	.01	.1221	.8440	1.0323	.0808	.8930	0.9760	.0378	.9440	1.0077	.0204	.9460	0.9819	.0040	.9450	0.9799
		.05	.1156	.8590	1.0206	.0795	.9000	0.9768	.0329	.9400	0.9950	.0217	.9320	0.9681	.0027	.9560	1.0533
		.10	.1250	.8440	1.0136	.0753	.9050	1.0200	.0376	.9310	0.9820	.0214	.9400	1.0049	.0080	.9450	0.9828
		.20	.1164	.8530	1.0150	.0759	.9000	0.9940	.0403	.9350	0.9989	.0178	.9410	0.9894	.0052	.9580	1.0157
	.65	.01	.3768	.1790	0.9792	.2498	.5060	1.0233	.1322	.8310	0.9955	.0596	.9270	0.9876	.0135	.9480	1.0014
		.05	.3677	.1700	1.0074	.2448	.5270	0.9782	.1249	.8260	0.9749	.0604	.9210	1.0037	.0021	.9440	0.9610
		.10	.3700	.1660	1.0439	.2456	.5260	1.0248	.1248	.8500	1.0019	.0659	.9230	1.0028	.0018	.9450	0.9874
		.20	.3799	.1570	1.0223	.2557	.4880	0.9879	.1258	.8490	1.0087	.0648	.9290	1.0449	.0173	.9440	0.9946
10,000	.45	.01	−.1211	.1510	0.9838	−.0794	.5020	0.9382	−.0417	.8140	1.0021	−.0222	.9240	1.0272	−.0014	.9460	0.9907
		.05	−.1201	.1510	0.9765	−.0810	.4700	1.0168	−.0399	.8320	1.0117	−.0183	.9190	0.9928	−.0018	.9440	0.9919
		.10	−.1200	.1490	0.9682	−.0813	.4550	1.0213	−.0398	.8330	1.0032	−.0199	.9100	0.9696	−.0028	.9550	1.0037
		.20	−.1179	.1730	0.9627	−.0793	.4890	0.9764	−.0411	.8190	0.9954	−.0199	.9330	1.0578	−.0029	.9550	1.0297
	.55	.01	.1195	.1470	1.0344	.0802	.4830	0.9700	.0383	.8320	1.0036	.0200	.9150	0.9717	.0058	.9470	0.9691
		.05	.1211	.1400	0.9898	.0800	.4870	0.9555	.0407	.8260	1.0384	.0193	.9200	1.0074	.0060	.9370	0.9844
		.10	.1221	.1420	0.9849	.0814	.4810	0.9800	.0373	.8600	1.0222	.0180	.9310	1.0053	.0010	.9610	1.0464
		.20	.1213	.1360	1.0191	.0786	.5080	1.0016	.0413	.8110	0.9906	.0214	.9160	0.9963	.0027	.9530	1.0121
	.65	.01	.3727	.0000	0.9855	.2499	.0000	0.9948	.1251	.1420	1.0078	.0617	.6740	0.9620	.0104	.9520	1.0347
		.05	.3735	.0000	1.0096	.2484	.0000	1.0109	.1270	.1300	0.9911	.0649	.6380	0.9782	.0145	.9310	1.0156
		.10	.3746	.0000	0.9988	.2493	.0000	1.0022	.1241	.1580	0.9729	.0624	.6740	1.0216	.0115	.9360	0.9955
		.20	.3706	.0000	1.0016	.2510	.0000	0.9942	.1266	.1250	1.0222	.0642	.6440	0.9656	.0115	.9450	1.0145

Table 6. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the unconditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	.0157	.9350	0.9918	.0090	.9530	1.0346	−.0021	.9520	1.0093	.0015	.9460	0.9795	−.0053	.9500	1.0045
		.05	.0081	.9410	1.0813	−.0010	.9320	0.9798	.0024	.9490	1.0050	−.0047	.9530	1.0270	.0029	.9510	0.9889
		.10	.0167	.9310	0.9873	−.0028	.9410	0.9611	.0050	.9450	0.9953	.0011	.9580	1.0166	.0016	.9460	0.9968
		.20	.0112	.9210	0.9989	.0079	.9400	0.9890	.0006	.9520	0.9672	−.0046	.9580	1.0020	−.0005	.9520	1.0285
	.55	.01	−.0308	.9310	0.9980	−.0031	.9350	0.9906	−.0006	.9440	0.9985	−.0093	.9470	0.9968	.0084	.9520	1.0232
		.05	−.0090	.9230	0.9695	−.0047	.9350	0.9788	−.0006	.9540	0.9978	−.0043	.9520	1.0365	−.0002	.9610	1.0240
		.10	−.0214	.9310	0.9734	−.0039	.9440	1.0123	−.0013	.9420	0.9833	.0019	.9450	0.9809	−.0045	.9600	1.0210
		.20	−.0278	.9360	0.9902	−.0112	.9460	1.0053	.0018	.9510	1.0056	.0049	.9560	1.0109	.0006	.9560	1.0235
	.65	.01	−.0345	.9310	1.0286	−.0049	.9470	1.0130	−.0013	.9360	0.9722	−.0071	.9500	1.0091	−.0009	.9490	0.9949
		.05	−.0444	.9320	1.0059	−.0098	.9400	0.9911	−.0001	.9510	1.0101	.0024	.9580	1.0171	.0028	.9420	0.9774
		.10	−.0578	.9190	0.9891	−.0149	.9470	1.0139	−.0062	.9590	0.9939	−.0091	.9430	0.9583	−.0054	.9530	0.9910
		.20	−.0361	.9460	1.0222	−.0051	.9430	1.0044	−.0044	.9600	1.0294	−.0080	.9510	1.0019	−.0048	.9410	0.9982
10,000	.45	.01	−.0046	.9280	0.9998	.0031	.9410	0.9762	.0024	.9530	1.0250	−.0022	.9550	1.0359	.0006	.9510	0.9851
		.05	−.0020	.9210	0.9806	.0005	.9500	1.0033	.0016	.9410	1.0022	.0021	.9380	0.9844	.0014	.9500	0.9966
		.10	.0044	.9200	0.9365	−.0017	.9520	1.0266	.0018	.9670	1.0462	−.0010	.9510	1.0170	.0018	.9510	1.0040
		.20	.0038	.9110	0.9531	.0012	.9510	0.9738	.0002	.9560	1.0014	−.0003	.9680	1.0423	−.0019	.9400	0.9699
	.55	.01	−.0043	.9150	0.9728	−.0009	.9350	0.9699	−.0019	.9360	0.9678	−.0000	.9510	0.9889	−.0002	.9410	0.9808
		.05	−.0034	.9320	0.9775	−.0007	.9320	0.9671	−.0003	.9480	1.0049	−.0004	.9550	1.0133	.0019	.9500	0.9753
		.10	−.0080	.9310	0.9924	−.0009	.9530	1.0127	−.0033	.9430	0.9760	.0003	.9510	0.9899	−.0005	.9450	0.9955
		.20	−.0069	.9210	0.9813	−.0028	.9460	1.0044	−.0010	.9340	0.9476	.0008	.9550	1.0331	.0001	.9490	1.0081
	.65	.01	−.0083	.9250	0.9643	−.0014	.9300	0.9458	−.0037	.9530	1.0423	−.0033	.9500	0.9952	−.0001	.9630	1.0809
		.05	−.0062	.9260	1.0011	−.0024	.9390	0.9908	−.0029	.9520	0.9845	.0002	.9410	0.9877	.0013	.9620	1.0461
		.10	−.0153	.9300	0.9918	−.0053	.9440	0.9832	−.0028	.9570	1.0140	−.0013	.9530	1.0042	.0008	.9570	1.0047
		.20	−.0134	.9030	0.9355	−.0032	.9370	0.9849	.0001	.9470	0.9967	−.0005	.9490	1.0010	−.0001	.9610	1.0555

Table 7. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	−.4303	.7100	1.3597	.0097	.9550	1.0372	−.0036	.9580	1.0172	.0022	.9450	0.9789	−.0053	.9470	1.0040
		.05	.0075	.9490	1.0825	−.0027	.9360	0.9812	.0025	.9520	0.9975	−.0043	.9550	1.0289	.0032	.9500	0.9899
		.10	.0135	.9310	0.9973	−.0021	.9410	0.9624	.0065	.9440	0.9964	.0014	.9570	1.0134	.0016	.9480	0.9982
		.20	.0027	.9290	0.9731	.0027	.9410	1.0048	.0003	.9540	0.9760	−.0048	.9470	0.9990	−.0006	.9540	1.0283
	.55	.01	.4050	.7500	1.2376	−.0039	.9360	0.9990	−.0015	.9470	1.0016	−.0096	.9490	0.9959	.0084	.9520	1.0231
		.05	−.0028	.9220	1.0031	−.0041	.9450	0.9927	−.0008	.9490	0.9984	−.0037	.9540	1.0433	−.0001	.9590	1.0239
		.10	−.0113	.9490	0.9894	−.0018	.9470	1.0198	−.0020	.9470	0.9814	.0025	.9460	0.9805	−.0044	.9600	1.0211
		.20	−.0087	.9450	1.0109	−.0080	.9490	1.0205	.0015	.9490	0.9990	.0056	.9590	1.0146	.0006	.9560	1.0236
	.65	.01	1.3336	.0800	1.3089	−.0060	.9480	1.0146	−.0027	.9440	0.9774	−.0076	.9520	1.0107	−.0010	.9540	0.9959
		.05	−.0381	.9430	1.6091	−.0101	.9370	0.9926	−.0012	.9540	1.0142	.0018	.9560	1.0148	.0028	.9420	0.9773
		.10	−.0316	.9310	1.0027	−.0114	.9460	1.0251	−.0070	.9590	0.9969	−.0089	.9340	0.9574	−.0052	.9520	0.9904
		.20	−.0004	.9480	1.0298	.0040	.9450	0.9902	−.0037	.9570	1.0298	−.0082	.9560	1.0068	−.0048	.9430	0.9975
10,000	.45	.01	.0027	.9240	0.9560	.0031	.9390	0.9819	−.0003	.9450	0.9913	−.0009	.9510	1.0202	.0030	.9410	0.9946
		.05	.0011	.9400	1.0213	.0005	.9490	1.0116	.0021	.9560	1.0079	.0020	.9420	0.9893	.0024	.9450	0.9938
		.10	.0004	.9220	0.9820	−.0026	.9540	1.0325	.0004	.9530	1.0206	.0008	.9430	0.9837	.0015	.9600	0.9878
		.20	.0006	.9380	0.9926	−.0005	.9510	0.9893	−.0019	.9550	1.0067	.0009	.9590	1.0582	.0005	.9640	1.0294
	.55	.01	−.0070	.9220	0.9893	−.0012	.9350	0.9709	−.0018	.9440	0.9918	−.0009	.9470	0.9815	.0018	.9470	0.9713
		.05	.0012	.9200	0.9515	−.0003	.9320	0.9704	.0005	.9520	0.9998	−.0012	.9570	1.0107	.0018	.9420	0.9900
		.10	−.0001	.9260	0.9837	−.0001	.9570	1.0173	−.0044	.9530	1.0407	−.0028	.9480	1.0010	−.0025	.9580	1.0482
		.20	.0029	.9290	0.9886	−.0011	.9430	1.0107	−.0001	.9410	0.9692	.0013	.9520	0.9884	−.0009	.9550	1.0200
	.65	.01	−.0105	.9220	1.1451	−.0026	.9300	0.9572	−.0044	.9580	1.0195	−.0029	.9530	0.9930	−.0022	.9540	1.0358
		.05	−.0074	.9280	0.9838	−.0025	.9390	0.9986	−.0011	.9500	0.9978	.0004	.9370	0.9773	.0014	.9540	1.0102
		.10	−.0014	.9400	0.9904	−.0034	.9410	0.9913	−.0043	.9500	1.0043	−.0017	.9490	1.0400	−.0009	.9530	0.9892
		.20	.0017	.9330	1.0127	.0023	.9470	0.9892	−.0024	.9480	0.9917	.0002	.9450	0.9932	−.0011	.9600	1.0095

Table 8. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the restricted conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	.0318	.9170	0.9730	.0099	.9470	1.0215	.0018	.9530	0.9893	−.0054	.9400	0.9693	−.0013	.9500	1.0134
		.05	.0099	.9390	1.0035	−.0011	.9380	0.9586	−.0096	.9410	0.9932	.0077	.9490	0.9912	.0024	.9500	0.9816
		.10	.0057	.9370	1.0108	−.0027	.9420	0.9728	.0047	.9580	1.0362	−.0029	.9520	1.0184	−.0033	.9430	0.9480
		.20	.0060	.9350	1.0079	.0050	.9530	1.0086	.0063	.9550	1.0681	.0051	.9530	1.0049	−.0008	.9450	0.9851
	.55	.01	−.0088	.9370	1.0423	−.0040	.9380	0.9916	−.0055	.9540	1.0004	−.0011	.9470	0.9818	−.0002	.9480	0.9880
		.05	−.0048	.9400	1.0348	−.0037	.9460	0.9922	−.0094	.9480	1.0012	.0016	.9430	0.9808	−.0014	.9550	1.0565
		.10	.0058	.9340	1.0123	−.0021	.9350	0.9904	−.0011	.9460	0.9986	.0031	.9510	1.0202	.0023	.9530	0.9840
		.20	−.0010	.9320	0.9809	−.0065	.9470	1.0168	−.0019	.9510	0.9913	−.0015	.9490	0.9833	.0009	.9550	1.0090
	.65	.01	−.0351	.9450	1.0301	−.0061	.9480	1.0107	.0008	.9450	0.9898	−.0086	.9530	1.0075	.0004	.9540	1.0022
		.05	−.0192	.9410	0.9847	−.0044	.9460	1.0003	−.0059	.9430	0.9849	−.0033	.9540	1.0015	−.0108	.9460	0.9629
		.10	−.0080	.9340	1.0162	−.0139	.9480	1.0160	−.0038	.9580	1.0093	−.0006	.9580	0.9895	−.0109	.9430	0.9965
		.20	.0016	.9460	0.9996	.0044	.9490	0.9849	.0044	.9550	1.0192	−.0029	.9600	1.0578	.0045	.9480	0.9885
10,000	.45	.01	−.0039	.9280	1.0269	.0033	.9340	0.9739	−.0005	.9470	0.9897	−.0005	.9540	1.0205	.0030	.9410	0.9939
		.05	−.0019	.9260	0.9895	.0006	.9510	1.0142	.0020	.9500	1.0154	.0019	.9400	0.9886	.0024	.9430	0.9934
		.10	−.0006	.9190	0.9415	−.0024	.9510	1.0333	.0003	.9520	1.0188	.0006	.9400	0.9867	.0015	.9580	0.9889
		.20	−.0004	.9370	0.9883	−.0005	.9550	1.0033	−.0019	.9510	1.0037	.0011	.9610	1.0530	.0006	.9620	1.0290
	.55	.01	−.0065	.9230	0.9828	−.0015	.9380	0.9802	−.0021	.9500	0.9876	−.0007	.9490	0.9789	.0019	.9450	0.9721
		.05	.0032	.9130	0.9552	.0003	.9380	0.9804	−.0000	.9490	1.0015	−.0011	.9590	1.0104	.0018	.9420	0.9893
		.10	.0008	.9370	0.9855	.0006	.9520	1.0206	−.0046	.9540	1.0307	−.0030	.9490	1.0022	−.0024	.9580	1.0489
		.20	.0041	.9360	0.9792	−.0007	.9490	1.0056	.0002	.9380	0.9725	.0012	.9510	0.9893	−.0009	.9560	1.0200
	.65	.01	−.0136	.9270	0.9887	−.0031	.9310	0.9522	−.0040	.9530	1.0141	−.0028	.9500	0.9954	−.0022	.9540	1.0352
		.05	−.0047	.9320	0.9980	−.0024	.9440	1.0046	−.0009	.9550	0.9956	.0006	.9360	0.9753	.0014	.9520	1.0110
		.10	.0016	.9370	0.9886	−.0028	.9460	0.9957	−.0038	.9460	0.9928	−.0018	.9510	1.0382	−.0009	.9510	0.9892
		.20	.0058	.9370	1.0049	.0029	.9460	0.9893	−.0022	.9510	0.9946	.0001	.9460	0.9941	−.0011	.9600	1.0092

Table 9. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the restricted conditional model and covariate Z with classification probabilities 0.90, P (Z = 2) = 0.1, sample size = 1000,00 and different values for the number of bootstrap samples m.

		class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
m	θ	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
	1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663	.0006	.9480	1.0044
	2	−.0705	.7450	0.9894	−.0423	.6770	0.9283	−.0137	.8760	0.9607	−.0041	.9320	0.9822	−.0009	.9540	1.0052
5	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	−.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	.07146	.7750	0.9870	.0427	.7090	0.9399	.014	.8780	0.9829	.0042	.9340	0.9690	.0006	.9460	1.0042
	2	−.0715	.7660	0.9929	−.0424	.6970	0.9406	−.0137	.8600	0.9685	−.0040	.9390	0.9861	−.0009	.9540	1.0056
10	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	.0710	.8010	1.0161	.0425	.7090	0.9360	.0138	.8880	0.9753	.0041	.9300	0.9646	.0006	.9480	1.0052
	2	−.0710	.8060	1.0222	−.0423	.6930	0.9399	−.0135	.8680	0.9626	−.0040	.9340	0.9828	−.0009	.9540	1.0061
20	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	0.0720	.7910	1.0152	.0426	.7120	0.9485	.0141	.8720	0.9826	.0042	.9330	0.9661	.0007	.9480	1.0045
	2	−0.0720	.7920	1.0208	−.0424	.7000	0.9512	−.0137	.8610	0.9692	−.0040	.9400	0.9835	−.0009	.9520	1.0055
40	3	0.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	−0.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994

Appendix B

- **Table 10 (Application) entropy R^2, classification probabilities, marginal probabilities:** This table shows the entropy $R2$, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the *rent benefit* variable takes information of 779 individuals into account and *marital status* variable of 3,011.
- **Table 11 (Application) proportions and marginal proportions:** The first block of tihs table represents the (pooled) marginal proportions of the variable *own/rent*. The second block represents the (pooled) proportions of the variable *own/rent* for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable *own/rent* for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.
- **Table 12 (Application) estimates of intercept and logit coefficient:** In this table, first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable *owning/renting* a house.

Table 10. Entropy R^2, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the rent benefit variable takes information of 779 individuals into account and marital status variable of 3,011.

			Unconditional model	Conditional model	Restricted conditional model
Entropy R^2			0.9334	0.9377	0.9380
	LISS	P(rent| LC rent)	0.8937	0.8938	0.9344
Classification	background	P(own| LC own)	0.9997	0.9997	0.9992
probability	BAG	P(rent| LC rent)	0.9501	0.9500	0.9496
	register	P(own| LC own)	0.9749	0.9749	0.9525
P(rent benefit)				0.3004	0.3004
P(married)			0.5284	0.5284	0.5284

Table 11. The first block represents the (pooled) marginal proportions of the variable own/rent. The second block represents the (pooled) proportions of the variable own/rent for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable own/rent for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
LISS background	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
Unconditional	0.6405	[0.6404; 0.6407]	0.3595	[0.3593; 0.3596]
Conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
Restricted conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
	P(own × rent benefit)		P(rent × rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
Unconditional	0.0028	[0.0023; 0.0034]	0.2950	[0.2944; 0.2955]
Conditional	0.0064	[−0.0263; 0.0392]	0.2914	[0.2587; 0.3241]
Restricted conditional	0.0000	-	0.2978	[0.2649; 0.3307]
	P(own × no rent benefit)		P(rent × no rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
Unconditional	0.0157	[0.0151; 0.0162]	0.6829	[0.6824; 0.6835]
Conditional	0.0159	[−0.0168; 0.0487]	0.6827	[0.6499; 0.7154]
Restricted conditional	0.0213	[−0.0116; 0.0542]	0.6773	[0.6444; 0.7102]

Table 12. The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable owning/renting a house.

	Intercept		Marriage	
	Estimate	95% CI	Estimate	95% CI
BAG register	2.4661	[2.2090; 2.7233]	−1.2331	[−1.3901; −1.0760]
LISS background	2.7620	[2.4896; 3.0343]	−1.3041	[−1.4678; −1.1405]
Unconditional	2.6869	[2.4251; 2.9487]	−1.3875	[−1.6493; −1.1257]
Conditional	2.7698	[2.5034; 3.0363]	−1.3982	[−1.6646; −1.1317]
Restricted conditional	2.7712	[2.5036; 3.0389]	−1.3817	[−1.6493; −1.1140]

6. References

André, S. and C. Dewilde. 2016. "Home Ownership and Support for Government Redistribution." *Comparative European Politics* 14: 319–348. Doi: http://dx.doi.org/10.1057/cep.2014.31.

Bakk, Z., D.L. Oberski, and J.K. Vermunt. 2016. "Relating Latent Class Membership to Continuous Distal Outcomes: Improving the LTB Approach and a Modified Three-Step Implementation." *Structural Equation Modeling: A Multidisciplinary Journal* 23: 278–289. Doi: http://dx.doi.org/10.1080/10705511.2015.1049698.

Bakker, B.F.M. 2009. Trek alle registers open! *Rede in verkorte vorm uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Methodologie van registers voor sociaalwetenschappelijk onderzoek bij de Faculteit der Sociale Wetenschappen van de Vrije Universiteit Amsterdam op 26 november 2009*. Available at: http://dare.ubvu.vu.nl/bitstream/handle/1871/15588/Oratie%20Bakker.pdf (accessed April 24, 2017).

Bakker, B.F.M. 2010. "Micro-Integration, State of the Art." *Paper presented at the joint UNECE-Eurostat expert group meeting on registered based censuses in The Hague, May 11, 2010*. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2010/wp.10.e.pdf (accessed April 24, 2017).

Bakker, B.F.M. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66: 8–17. Doi: http://dx.doi.org/10.1111/j.14679574.2011.00504.x.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Error* (Vol. 571). Hoboken, New Jersey: John Wiley & Sons.

De Waal, T. 2016. "Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys." *Statistical Journal of the IAOS* 32: 231–243. Doi: http://dx.doi.org//10.3233/SJI-150950.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation* (Vol. 563). John Wiley & Sons.

De Waal, T., J. Pannekoek, and S. Scholtus. 2012. "The Editing of Statistical Data: Methods and Techniques for the Efficient Detection and Correction of Errors and Missing Values." *Wiley Interdisciplinary Reviews: Computational Statistics* 4: 204–210. Doi: http://dx.doi.org/10.1002/wics.1194.

Dewilde, C. and P.D. Decker. 2016. "Changing Inequalities in Housing Outcomes Across Western Europe." *Housing, Theory and Society* 33: 121–161. Doi: http://dx.doi.org/10.1080/14036096.2015.1109545.

Dias, J.G. and J.K. Vermunt. 2008. "A Bootstrap-Based Aggregate Classifier for Model-Based Clustering." *Computational Statistics* 23: 643–659. Doi: http://dx.doi.org/10.1007/s00180-007-0103-7.

Forcina, A. 2008. "Identifiability of Extended Latent Class Models with Individual Covariates." *Computational Statistics & Data Analysis* 52: 5263–5268. Doi: http://dx.doi.org/10.1016/j.csda.2008.04.030.

Geerdinck, M., M. Goedhuys-van der Linden, E. Hoogbruin, A. De Rijk, N. Sluiter, and C. Verkleij. 2014. *Monitor Kwaliteit Stelsel van Basisregistraties: Nulmeting van de Kwaliteit van Basisregistraties in Samenhang, 2014* (13114th ed.). Henri Faas-dreef 312, 2492 JP Den Haag: Centraal Bureau voor de Statistiek. Available at:

https://www.cbs.nl/-/media/pdf/2016/50/monitor-kwaliteit-stelsel-van-basisregistraties.pdf (accessed April 25, 2017).

Groen, J.A. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28: 173–198.

Guarnera, U. and R. Varriale. 2016. "Estimation from Contaminated Multi-Source Data Based on Latent Class Models." *Statistical Journal of the IAOS* 32: 537–544. Doi: dx.doi.org//10.3233/SJI-150951.

Jörgren, F., R. Johansson, L. Damber, and G. Lindmark. 2010. "Risk Factors of Rectal Cancer Local Recurrence: Population-Based Survey and Validation of the Swedish Rectal Cancer Registry." *Colorectal Disease* 12: 977–986. Doi: http://dx.doi.org/10.1111/j.1463-1318.2009.01930.x.

Kim, H.J., L.H. Cox, A.F. Karr, J.P. Reiter, and Q. Wang. 2015. "Simultaneous Edit-Imputation for Continuous Microdata." *Journal of the American Statistical Association* 110: 987–999. Doi: http://dx.doi.org/10.1080/01621459.2015.1040881.

Lersch, P.M. and C. Dewilde. 2015. "Employment Insecurity and First-Time Homeownership: Evidence from Twenty-Two European Countries." *Environment and Planning A* 47: 607–624. Doi: http://dx.doi.org//10.1068/a130358p.

Manrique-Vallier, D. and J.P. Reiter. 2013. "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros." *Survey Methodology* 40: 125–134. Available at: https://ecommons.cornell.edu/handle/1813/34889 (accessed April 25, 2017).

Manrique-Vallier, D. and J.P. Reiter. 2016. "Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data." *Journal of the American Statistical Association*. Doi: http://dx.doi.org/10.1080/01621459.2016.1231612.

Mulder, C.H. 2006. "Home-Ownership and Family Formation." *Journal of Housing and the Built Environment* 21: 281–298. Doi: http://dx.doi.org/10.1007/s10901-006-9050-9.

Ness, A.R. 2004. "The Avon Longitudinal Study of Parents and Children (ALSPAC)- a Resource for the Study of the Environmental Determinants of Childhood Obesity." *European Journal of Endocrinology 151*(Suppl 3): U141–U149. Doi: http://dx.doi.org//10.1530/eje.0.151U141.

Oberski, D.L. 2015. "Total Survey Error in Practice." In *Total Survey Error*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, and B. West. New York: Wiley.

Pavlopoulos, D. and J. Vermunt. 2015. "Measuring Temporary Employment. Do Survey or Register Tell the Truth?" *Survey Methodology* 41: 197–214. Available at: http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14151-eng.pdf (accessed April 25, 2017).

R Core Team. 2014. "R: A Language and Environment for Statistical Computing [Computer software manual]." Vienna, Austria. Available at: http://www.R-project.org/ (accessed October 13, 2017).

Robertsson, O., M. Dunbar, K. Knutson, S. Lewold, and L. Lidgren. 1999. "Validation of the Swedish Knee Arthroplasty Register: A Postal Survey Regarding 30,376 Knees Operated on Between 1975 and 1995." *Acta Orthopaedica Scandinavica* 70: 467–472. Doi: http://dx.doi.org/10.3109/17453679909000982.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys* (Vol. 81). John Wiley & Sons. Doi: http://dx.doi.org//10.1002/9780470316696.

Scherpenzeel, A. 2011. "Data Collection in a Probability-Based Internet Panel: How the LISS Panel was Built and How it can be Used." *Bulletin of Sociological Methodology/Bulletin de Methodologie Sociologique* 109: 56–61. Doi: http://dx.doi.org//10.1177/0759106310387713.

Scholtus, S. 2009. "Automatic Detection of Simple Typing Errors in Numerical Data with Balance Edits." *Statistics Netherlands Discussion Paper* (09046). Available at: https://www.cbs.nl/-/media/imported/documents/2009/48/2009-46-x10-pub.pdf (accessed April 25, 2017).

Scholtus, S. 2011. "Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data." *Journal of Official Statistics* 27: 467–490.

Scholtus, S. and B.F.M. Bakker. 2013. "Estimating the Validity of Administrative and Survey Variables through Structural Equation Modeling: A Simulation Study on Robustness." *Statistics Netherlands Discussion Paper*. Available at: https://www.cbs.nl/-/media/imported/documents/2013/12/2013-02-x10-pub.pdf (accessed April 25, 2017).

Schrijvers, C.T.M., K. Stronks, D.H. van de Mheen, J.-W. W. Coebergh, and J.P. Mackenbach. 1994. "Validation of Cancer Prevalence Data from a Postal Survey by Comparison with Cancer Registry Records." *American Journal of Epidemiology* 139: 408–414. Doi: https://doi.org/10.1093/oxfordjournals.aje.a117013.

Schulte Nordholt, E., J. Van Zeijl, and L. Hoeksma. 2014. "Dutch Census 2011, Analysis and Methodology." *Statistics Netherlands*. Available at: https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf (accessed April 25, 2017).

Si, Y. and J.P. Reiter. 2013. "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys." *Journal of Educational and Behavioral Statistics* 38: 499–521. Doi: dx.doi.org//10.3102/1076998613480394.

Tempelman, C. 2007. *Imputation of Restricted Data: Applications to Business surveys* (Doctoral dissertation, Rijksuniversiteit Groningen). Available at: https://www.cbs.nl/-/media/imported/documents/2007/05/2007-i76-pub.pdf (accessed April 25, 2017).

Turner, C.F., T.K. Smith, L.K. Fitterman, T. Reilly, K. Pate, M.B. Witt, and B.H. Forsyth. 1997. "The Quality of Health Data Obtained in a New Survey of Elderly Americans: A Validation Study of the Proposed Medicare Beneficiary Health Status Registry (mbhsr)." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 52B: S49–S58. Doi: http://dx.doi.org//10.1093/geronb/52B.1.S49.

Understanding Society. 2016. "Understanding Society: Innovation Panel, Waves 1–7, 2008–2014 [data collection]. 6th edition [Computer software manual]. UK Data Service. Doi: 10.5255/UKDA-SN-6849-7.

University of London. Institute of Education. Centre for Longitudinal Studies, Millennium Cohort Study: First Survey, 2001–2003 [computer file]. 6th edition. Colchester, Essex: UK Data Archive [distributor], SN: 4683. (2007, March). Available at: http://dx.doi.org/10.5255/UKDA-SN-4683-1.

Van der Palm, D.W., L.A. Van der Ark, and J.K. Vermunt. 2016. "Divisive Latent Class Modeling as a Density Estimation Method for Categorical Data." *Journal of Classification* 1–21. Doi: http://dx.doi.org/10.1007/s00357-016-9195-5.

Van der Vaart, W. and T. Glasner. 2007. "Applying a Timeline as a Recall Aid in a Telephone Survey: a Record Check Study." *Applied Cognitive Psychology* 21: 227–238. Doi: http://dx.doi.org/10.1002/acp.1338.

Vermunt, J.K. and J. Magidson. 2004. "Latent Class Analysis." *The Sage Encyclopedia of Social Sciences Research Methods* 549–553. Available at: http://members.home.nl/jeroenvermunt/ermss2004a.pdf (accessed April 25, 2017).

Vermunt, J.K. and J. Magidson. 2013a. Latent GOLD 5.0 Up-grade Manual [Computer software manual]. Belmont, MA. Available at: https://www.statisticalinnovations.com/wp-content/uploads/LG5manual.pdf (accessed April 25, 2017).

Vermunt, J.K. and J. Magidson. 2013b. "Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax." *Statistical Innovations Inc., Belmont, MA*. Available at: https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf (accessed April 25, 2017).

Vermunt, J.K., J.R. Van Ginkel, L.A. Van Der Ark, and K. Sijtsma. 2008. "Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis." *Sociological Methodology* 38: 369–397. Doi: http://dx.doi.org/10.1111/j.1467-9531.2008.00202.x.

Vink, G. and S. van Buuren. 2014. "Pooling Multiple Imputations When the sample Happens to be the Population." *arXiv preprint arXiv:1409.8542*. Available at: https://arxiv.org/abs/1409.8542.

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

Zhang, L.-C. and J. Pannekoek. 2015. "Optimal Adjustments for Inconsistency in Imputed Data." *Survey Methodology* 41: 127–144. Available at: http://www.statcan.gc.ca/pub/12-001-x/12-001-x2015001-eng.pdf (accessed April 25, 2017).

Received July 2016
Revised April 2017
Accepted May 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 963–978, http://dx.doi.org/10.1515/JOS-2017-0045

How to Obtain Valid Inference under Unit Nonresponse?

Laura Boeschoten[1], *Gerko Vink*[2], *and Joop J.C.M. Hox*[2]

Weighting methods are commonly used in situations of unit nonresponse with linked register data. However, several arguments in terms of valid inference and practical usability can be made against the use of weighting methods in these situations. Imputation methods such as sample and mass imputation may be suitable alternatives, as they lead to valid inference in situations of item nonresponse and have some practical advantages. In a simulation study, sample and mass imputation were compared to traditional weighting when dealing with unit nonresponse in linked register data. Methods were compared on their bias and coverage in different scenarios. Both, sample and mass imputation, had better coverage than traditional weighting in all scenarios.

Imputation methods can therefore be recommended over weighting as they also have practical advantages, such as that estimates outside the observed data distribution can be created and that many auxiliary variables can be taken into account. The use of sample or mass imputation depends on the specific data structure.

Key words: Weighting; mass imputation; sample imputation; coverage.

1. Introduction

Missing data form a ubiquitous source of problems in survey research. A common research scenario occurs when respondents that are sampled from the population cannot be contacted, or when they are reluctant to conform to the survey. If no analysable information about the respondent is collected, we deem it unit nonresponse. In such a scenario, we can distinguish between two missing data problems. The first problem is that, when sampling from the population, not all units from the population are recorded (which is the usual process of sampling producing missing data by design). The second problem is that the sample is found to be incomplete. The severity of these problems is related to the probability each data point has of being missing.

The mechanism that governs these probabilities is called the missing data mechanism (Rubin 1976). To describe these mechanisms, we assume to have a data set consisting of an incomplete target variable Y and a fully observed covariate X. The incomplete target variable Y has two parts: an observed part Y_{obs} and a missing part Y_{mis}. An indicator variable R can be defined that scores a 0 when Y is missing and a 1 when Y is observed.

[1] Tilburg School of Social and Behavioral Sciences, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: L.Boeschoten@tilburguniversity.edu
[2] Department of Methodology & Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: G.Vink@uu.nl and J.Hox@uu.nl
Acknowledgments: The authors would like to thank the anonymous associate editor and three referees for their constructive feedback on an earlier version of this manuscript. Furthermore, the authors would like to thank Barry Schouten and Sander Scholtus for their valuable input.

If the data are Missing Completely At Random (MCAR, Rubin 1976), the response probability for the respondents and nonrespondents is equal. This can be formally defined as:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0) \tag{1}$$

"An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck" (Van Buuren 2012, 7). If the data are Missing At Random (MAR, Rubin 1976), the distribution of the missing values is related to other observed values, formally defined:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0|Y_{obs}, X) \tag{2}$$

"For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and if we can assume MCAR within the type of surface, then the data are MAR" (Van Buuren 2012, 7). If the distribution of the missing values relates to unobserved values, it is called Missing Not At Random (MNAR, Rubin 1976), formally defined:

$$P(R = 0|Y_{obs}, Y_{mis}, X) = P(R = 0|Y_{obs}, Y_{mis}, X) \tag{3}$$

"For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted" (Van Buuren 2012, 7).

Sometimes register data is available with information about the characteristics of the respondents and the nonrespondents that can be linked to the survey data (Bethlehem et al. 2011, 211). If there is a relationship between the selection mechanism and the survey variables, the estimators will systematically over- or under-represent the population characteristics. Such deviations can be corrected by weighting the observed data to conform to the known population parameters. If done properly, both distinct missing data problems can in theory be solved. However, there are several arguments against the use of weighting techniques to handle nonresponse. We list them in no particular order:

1. Weighting ignores the uncertainty about the missing data. This may result in too little variation about the estimates (Bethlehem et al. 2011, 184).
2. Weighting methods cannot create estimates that lie outside the observed data distribution. Although some researchers might view this as an advantage of weighting and would worry when a method could yield estimates outside the observed data distribution, an example given by Rubin illustrates when this could be problematic: "Consider dealing with censored data by weighting – data beyond or approaching the censoring point have zero or very small probabilities of being observed, and so either cannot be dealt with by weighting or imply a few observations with dominant weights. Weighting by inverse probabilities cannot create estimates outside the convex hull of the observed data, and estimates involving weights near the boundary have extremely large variance" (Rubin 1996, 486).

3. Uncertainty about the weights is ignored when weights are estimated from the data and thereby treated as fixed, given that the data conform to sampling variance. When taking additional measures, such as combining jackknife procedures with calibration, or by using design based analysis, weights can be treated as random.
4. Weighting has difficulties with handling large numbers of auxiliary variables, which are potentially needed to make the nonresponse ignorable (Rubin 1987, 155). Additional measures should then be taken, such as dimension reduction or propensity score estimation.
5. Weighting can have difficulties with creating sensible weights when more auxiliary information is incorporated. As a result, it is possible that the score on a target variable of an individual is used to represent a large group in the population. An illustrative example from the United States of America 2016 presidential elections show how one man heavily influenced the outcome of a poll due to extreme weights being given to his demographic category (Cohn 2016).
6. Some weighting methods cannot handle continuous variables.
7. Weighting cannot handle partial response. It is an all or nothing approach and may thereby discard valuable information (Van Buuren 2012, 22).

Because of arguments 1 and 3, we expect weighting to create too little variance and therefore to yield invalid inference (with confidence validity as defined by Rubin (1996)). We expect multiple imputation (MI) to be a good alternative method to correct for unit nonresponse, since it takes sampling variability as well as uncertainty due to missing values into account (Rubin 1987, 76). Furthermore, with MI there is no limit to the use of auxiliary information: continuous variables or the number of variables are less likely to pose problems, as the likelihood of the observed data given the unobserved data is taken into account. In cases of large numbers of variables or nonlinear associations, principal component analysis can be used (Howard 2012). In addition, item and unit non-response can be handled simultaneously with MI.

The goal of this article is to investigate whether MI is a suitable alternative for weighting when correcting for unit nonresponse. In this article, we distinguish between sample and mass imputation. With sample imputation, both item and unit nonresponse (occuring both in the sample) can be imputed. If the sample is a simple random sample without replacement (SRS_{WOR}) auxiliary information is only needed for the sample. However, sometimes registers with information about the whole population can be linked on a unit level to sample data sets. This is for example the case at Statistics Netherlands where complete population registers were used in the 2011 Dutch census (Schulte Nordholt et al. 2014). If this is the case, the nonsampled units can be imputed as well (besides the item and unit nonresponse within the sample). Mass imputation can then be applied with SRS_{WOR} or complex samples.

Our definition of mass imputation should not be confused with the approach of Zhou et al. (2016), who generate a synthetic data set based on known population totals. A benefit of mass imputation is that every source of (linked) auxiliary information can be used for imputation. This means that a MNAR missing mechanism can become MAR, leading to more efficient estimation of (population) parameters.

We investigate the performance of weighting and both sample and mass imputation. As a reference, we also investigate complete case analysis (CCA), where no correction for unit nonresponse is made. With performance, summarized as 'valid inference' in the title, we mean obtaining unbiased parameter estimates and unbiased variance estimates.

2. Methodology

In this article, we distinguish between multiple auxiliary variables $\mathbf{X}$ and a single target variable y, which we assume to be normally distributed with mean μ and variance σ^2. If we would take a SRS_{WOR}, the estimate of the sample mean of a target variable y is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{4}$$

where y_i is the observation on the i^{th} sampled unit with $i = 1, \ldots, n$, where n is the sample size. The estimate of the variance of the mean is:

$$\text{VAR}(\hat{\mu}) = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{\mu})^2 \frac{1}{n}\left(1 - \frac{n}{N}\right), \tag{5}$$

where N is the size of the (finite) population. This is how μ and $\text{VAR}(\mu)$ are estimated when the sample is completely observed. We will now discuss different methods to estimate these parameters in case of unit nonresponse.

2.1. Complete Case Analysis

When CCA is applied, nonrespondents are completely removed from the sample. μ and $\text{VAR}(\mu)$ are estimated with the same equations used for a completely observed sample, as in Equations 4 and 5. However, with unit nonresponse, not all values in y are observed, and only the observed values in y are used to estimate μ and $\text{VAR}(\mu)$ of the target variables:

$$\hat{\mu} = \frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}} y_{obs_i}, \tag{6}$$

$$\text{VAR}(\hat{\mu}) = \frac{1}{n_{obs}-1}\sum_{i=1}^{n_{obs}} (y_{obs_i} - \hat{\mu}_{obs})^2 \frac{1}{n_{obs}}\left(1 - \frac{n_{obs}}{N}\right). \tag{7}$$

2.2. Weighting

The weighted mean of a target variable is defined as

$$\hat{\mu} = \frac{\sum_{i=1}^{n_{obs}} w_i y_{obs_i}}{\sum_{i=1}^{n_{obs}} w_i} \tag{8}$$

where w_i is the weight corresponding to the i^{th} observation (Biemer and Christ 2008, 318) and μ is a vector quantity (and is so throughout the remainder of the article). The weights, w_i, can be estimated with different methods, such as poststratification, linear weighting,

multiplicative weighting and propensity weighting. A full description of how to apply the different methods can be found in Chapter 8 of Bethlehem et al. (2011). De Waal et al. (2011, 237–244) show that under certain conditions, linear weighting and mass imputation yield the same estimate. Therefore, it would be interesting to use this method to estimate the weights, and investigate whether these methods also yield the same inference. For this reason, we use linear weighting to estimate w_i.

Linear weighting is a calibration method, and is thoroughly discussed by, among others, Deville and Särndal (1992) and Särndal et al. (1992). When estimating weights, it is important to note first that these weights (w_i) consist of two parts:

$$w_i = d_i \delta_i, \tag{9}$$

where d_i are the sampling design weights. For a SRS_{WOR}, N and n are fixed numbers, d_i is constant and does not need to be estimated:

$$d_i = N/n. \tag{10}$$

δ_i is the adjustment factor. Our goal is to find a δ_i which makes w_i as close as possible to d_i, while respecting the calibration equation

$$\sum_{i=1}^{n_{obs}} w_i \mathbf{X}_i = \mathbf{t}_{\mathbf{X}}, \tag{11}$$

where $\mathbf{X}$ represents the auxiliary variables and $\mathbf{t}_{\mathbf{X}}$ are the population totals of $\mathbf{X}$. Minimizing the function

$$\sum_{i=1}^{n_{obs}} (w_i - d_i)^2 / d_i \tag{12}$$

leads to what is also known as linear weighting, which is a special case of calibration. We derive new weights here that modify as little as possible to the original sampling design weights d_i by minimizing the conditional value of the distance, given the realized observed sample n_{obs}. This leads to the calibrated weight

$$w_i = d_i(1 + \mathbf{X}_i' \lambda) \tag{13}$$

where λ is a vector of Lagrange multipliers determined from Equation 12:

$$\lambda = \mathbf{T}_{n_{obs}}^{-1} (\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}\pi}). \tag{14}$$

The inverse of $\mathbf{T}_{n_{obs}}$ is

$$\mathbf{T}_{n_{obs}}^{-1} = \left(\sum d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \tag{15}$$

and $\hat{\mathbf{t}}_{\mathbf{X}\pi}$ is the Horvitz-Thompson (Horvitz and Thompson 1952) estimator for $\mathbf{X}$:

$$\hat{\mathbf{t}}_{\mathbf{X}\pi} = \sum_{i=1}^{n_{obs}} d_i \mathbf{X}_{\mathbf{i}} \tag{16}$$

(Deville and Särndal 1992). The variance of a weighted mean can be approximated with methods such as Taylor linearization or Jacknife resampling (Stapleton 2008, 355). We

use Taylor linearization and we assume for convenience that there is a vector of constants γ, such that $\gamma' \mathbf{X}_i = 1$ for all i. In that case, $\sum_{i=1}^{n_{obs}} w_i = N$. Then, the variance of a weighted mean can be approximated by:

$$\mathrm{VAR}(\hat{\mu}) = \frac{1}{N^2} \sum_{i=1}^{n_{obs}} \sum_{h=1, h \neq i}^{n_{obs}} \frac{\pi_{ih} - \pi_i \pi_h}{\pi_{ih}} \left(\delta_i \frac{e_i}{\pi_i} \right) \left(\delta_h \frac{e_h}{\pi_h} \right) \tag{17}$$

where π_i and π_h are the first order and π_{ih} the corresponding second order inclusion probabilities of observations i and h, and e_i (and e_h) are defined as:

$$e_i = y_i - \mathbf{X}_i' \mathbf{T}_{n_{obs}}^{-1} \sum_{l=1}^{n_{obs}} \mathbf{X}_l y_l d_l \tag{18}$$

(Särndal et al. 1992, 225–236).

2.3. Sample Imputation

With MI, each missing datapoint is imputed $m \geq 2$ times, resulting in m completed data sets. At least two imputations are needed to reflect the uncertainty about the imputations, although performing more imputations is often advisable. The m data sets can then be analyzed by standard procedures and the analyses combined into a single inference. A clear introduction to multiple imputation and different methods to impute the missing datapoints is given in Van Buuren (2012, Chapter 2).

With sample imputation, we only impute the nonrespondents in the sample. Because the imputation theory aims at inference about the population, sampling uncertainty is taken into account and we can use the standard rules for pooling.

The pooled estimate of μ is obtained by

$$\bar{\mu} = \frac{1}{m} \sum_{j=1}^{m} \hat{\mu}_j \,, \tag{19}$$

where m is the number of imputations with $j = 1, \ldots, m$ and $\hat{\mu}_j$ is the $\hat{\mu}$ of the j^{th} imputed sample. $\overline{\mathrm{VAR}(\hat{\mu})}$ consists of multiple components (we therefore name it $\overline{\mathrm{VAR}(\hat{\mu})}_{\text{total}}$) and is estimated

$$\overline{\mathrm{VAR}(\hat{\mu})}_{\text{total}} = \overline{\mathrm{VAR}(\hat{\mu})}_{\text{within}} + \mathrm{VAR}(\hat{\mu})_{\text{between}} + \frac{\mathrm{VAR}(\hat{\mu})_{\text{between}}}{m}, \tag{20}$$

where $\overline{\mathrm{VAR}(\hat{\mu})}_{\text{within}}$ is the within imputation variance and $\mathrm{VAR}(\hat{\mu})_{\text{between}}$ is the between imputation variance. $\overline{\mathrm{VAR}(\hat{\mu})}_{\text{within}}$ is calculated by

$$\overline{\mathrm{VAR}(\hat{\mu})}_{\text{within}} = \frac{1}{m} \sum_{j=1}^{m} \mathrm{VAR}(\hat{\mu})_{\text{within}_j} \tag{21}$$

and $\mathrm{VAR}(\hat{\mu})_{\text{between}}$ is calculated by

$$\mathrm{VAR}(\hat{\mu})_{\text{between}} = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\mu}_j - \bar{\mu})(\hat{\mu}_j - \bar{\mu})'. \tag{22}$$

2.4. Mass Imputation

With mass imputation, the estimate of μ is also obtained by Equation 19, although $\hat{\mu}_j$ now corresponds to the j^{th} imputed version of the population instead of the the j^{th} imputed sample.

Because we impute the population, there is no variance due to sampling. Therefore, $\overline{\text{VAR}(\hat{\mu})}_{\text{within}} = 0$ and we can adjust Equation 20 to

$$\overline{\text{VAR}(\hat{\mu})}_{\text{total}} = \text{VAR}(\hat{\mu})_{\text{between}} + \frac{\text{VAR}(\hat{\mu})_{\text{between}}}{m}. \tag{23}$$

For a thorough description of making multiply imputed inference when sampling variance is not of interest see Vink and Van Buuren (2014).

3. Simulation Approach

To empirically evaluate the performance of the different analysis methods, we conducted a simulation study using R (R Core Team 2015, version 3.2.2). The properties we manipulate in the simulation design can be summarized as follows:

- The correlation between the auxiliary variables and the target variables: 0.30; 0.50.
- The amount of missingness: 25%; 50%.
- The missingness mechanism: MCAR; left-tailed MAR.
- The analysis method: CCA; lineair weighting (calibration); Bayesian normal linear imputation of the sample; Bayesian normal linear imputation of the population.

We now discuss the properties of the simulation design in more detail.

3.1. The Correlation Structure

We start by creating a large but finite population of 100,000 units with two auxiliary (X_1 and X_2) and two target variables (Y_1 and Y_2). The population data is multivariate normally distributed with $\boldsymbol{\mu}$ and Σ:

$$\begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{pmatrix} = MVN(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu}$ is:

$$\boldsymbol{\mu} = \begin{matrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{matrix} \begin{pmatrix} 3 \\ 2 \\ 0 \\ 170 \end{pmatrix}$$

and Σ is either:

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \end{array} \\ \begin{pmatrix} 1.00 & 0.08 & 1.34 & 1.90 \\ 0.08 & 0.25 & 0.67 & 0.95 \\ 1.34 & 0.67 & 20.00 & 4.24 \\ 1.90 & 0.95 & 4.24 & 40.00 \end{pmatrix} \end{array}$$

when the correlations between the target variables and the auxiliary variables are 0.30, and

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \end{array} \\ \begin{pmatrix} 1.00 & 0.08 & 2.24 & 3.16 \\ 0.08 & 0.25 & 1.12 & 1.58 \\ 2.24 & 1.12 & 20.00 & 4.24 \\ 3.16 & 1.58 & 4.24 & 40.00 \end{pmatrix} \end{array}$$

when the correlations between the target variables and the auxiliary variables are 0.50. The target variables X_1 and X_2 are transformed into categorical variables with respectively six and four categories, because auxiliary register information is in practice often categorical.

3.2. *The Amount of Missingness and the Missingness Mechanism*

From the population of size 100,000, a random sample of size 5,000 is drawn. In each sample, either 25% or 50% missingness is induced in the Y_1 and Y_2 variables.

The missingness in the target variables follow MCAR or left-tailed MAR mechanisms conform the procedure described by Van Buuren (2012, 63). With a left-tailed MAR mechanism, the probability of having missing values in the target variables is larger for smaller values on the auxiliary variables. For example, consider the number of employees of a company to be the auxiliary variable on which the missingness depends and working conditions of the company as target variables. In this situation, it is likely that more missing values are found at the companies with fewer employees. The first reason for this is that smaller companies are often less well organized. However, researchers are also probably more interested in larger companies, and are more likely to re-contact these in cases of nonresponse. If you sort companies on an axis with number of employees, you find more missing values on the left side of this axis, where the smaller companies are found.

3.3. *The Analysis Method*

We estimate $\hat{\mu}$ and VAR($\hat{\mu}$) of the target variables by making use of CCA, weighting, sample imputation and mass imputation. There are slight differences between the simulation setup within the different methods. For CCA, 96.25% or 97.50% of the 100,000 population values could be deleted directly from the target variables using MAR or MCAR to come to a sample of 5,000 with 25% or 50% missing values. The estimates of the incomplete sample can be compared directly to the population values.

Table 1. Smallest and largest adjustment factor per simulated condition.

		MCAR		MARleft	
cor.	% mis	min	max	min	max
0.3	25	1.2305	1.4511	0.9682	4.2467
	50	1.7472	2.3080	0.9419	13.8479
0.5	25	1.2298	1.4513	0.9646	4.2426
	50	1.7454	2.3128	0.9273	13.4502

For weighting, we first select randomly 5,000 cases from the population. Next, we create unit missingness following one of the missingness mechanisms. We weight the respondents to the total sample using the population totals. Weights are calculated using the survey package (Lumley 2014, version 3.30-3) in R (R Core Team 2015, version 3.2.2) with the `calibrate()` function. We evaluate the performance of weighting by comparing the results of the weighted sample to the population values. The design weights are $d_i = N/n = 100{,}000/5{,}000 = 20$. The adjustment factors δ_i can be found in Table 1, which can be used to compute the weights $w_i = d_i\delta_i$.

We are aware that some of the correction weights are considered large and that weighted estimates may be inefficient in such scenarios. An option would be to trim the weights to predefined boundaries. However, by not trimming the weights, we are able to investigate the performance of the method itself and its default options to other methods and their default options.

For sample imputation, we also 5,000 cases from the population and create unit missingness in the sample. Next, we multiply impute the sample and compare the results of the imputed sample to the population results.

For mass imputation, we can directly delete 96.25% or 97.50% of the values of the target variables and multiply impute the population. The results of the imputed population are compared to the original population results. Both sample and mass imputations are executed with `mice` (Van Buuren and Groothuis-Oudshoorn 2011) in `R` (R Core Team 2015) using Bayesian normal linear imputation (`mice.impute.norm()`) as the imputation method with five imputations and five iterations for the algorithm to converge.

3.4. *Performance Measures*

We estimate $\hat{\mu}$ and VAR($\hat{\mu}$) by using the previously discussed methods and replicate this procedure 1,000 times. In each replication, we investigate these estimates by looking at two performance measures. First, we look at the bias of $\hat{\mu}$ of the two target variables. This bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the $\hat{\mu}$'s of the two target variables over all replications.

3.5. *Expectations*

When CCA is applied and the missingness is MCAR, the probability of being missing is equal for every unit in the sample. Therefore, we do not expect biased estimates of $\hat{\mu}$.

However, with MAR, the probability of being missing is not equal for every unit, and we do expect bias. Since parameter uncertainty and uncertainty about the missing values is not taken into account when estimating the variance of the mean, we also expect undercoverage with MAR.

When weighting is applied, we expect unbiased estimates of $\hat{\mu}$ under both MCAR and MAR. The variance estimate takes the weights and parameter uncertainty into account, but not the uncertainty about the missing values. Therefore, we expect an estimate of the variance of the mean that is a bit too small, resulting in undercoverage under MAR.

For sample imputation we expect unbiased estimates and adequate coverage under both MCAR and MAR.

For mass imputation, we also expect unbiased estimates and adequate coverage under both MCAR and MAR.

4. Results

The simulation results are depicted in Table 2. Note that the results for CCA in terms of coverage and confidence interval width with correlation 0.30 and 0.50 look identical under MCAR. Small differences in the results were found, but these occur after the fourth decimal.

4.1. The Missingness Mechanism

The methods that aim to correct for the nonresponse show equivalent bias and coverage patterns under MCAR and left-tailed MAR missingness mechanisms. Naturally, the loss of observed information results in larger confidence interval widths under left tailed MAR missingness than under MCAR missingness mechanisms. CCA is unable to handle the estimation under left-tailed MAR missingness and yields large bias, zero coverage and confidence intervals that are, as expected, equally wide to those under MCAR.

4.2. The Correlation Structure

Larger correlations are often beneficial when solving incomplete data problems because the correlations give strong direction to the estimation procedure. This is clearly visible in all methods that aim to solve the missingness problem as confidence intervals tend to become smaller when the correlation between the target variables and the linked register data increases. Interestingly, the coverage rates for weighting are negatively impacted under large correlations. In this specific situation the bias remains roughly the same as under low-correlation simulations, while the confidence interval widths decrease. As a result, the simulations for weighting demonstrate lower coverage of the population mean.

4.3. The Amount of Missingness

In general, it can be said that when amounts of missingness become larger, incomplete data problems become more difficult. More specifically, the probability that you deal with a MNAR mechanism increases. None of the methods seem negatively impacted by the increased amount of missingness, when compared to the results under less missingness. However, the confidence intervals naturally tend to become wider as there is less information about the observed data.

Table 2. Simulation results. Depicted are the bias of the mean of Y_1 and Y_2, coverage of the 95% confidence interval and width of the 95% confidence interval for the four methods under varying simulation conditions.

				MCAR			MARleft		
Method	Correlation	% mis	Y	bias	coverage	CI width	bias	coverage	CI width
CCA	0.3	25	1	−0.0003	0.9620	0.2872	−1.1480	0.0000	0.2866
			2	0.0020	0.9580	0.4049	−1.6535	0.0000	0.4060
		50	1	−0.0003	0.9590	0.3516	−1.1728	0.0000	0.3479
			2	−0.0000	0.9620	0.4957	−1.6889	0.0000	0.4930
	0.5	25	1	−0.0002	0.9620	0.2872	−1.9291	0.0000	0.2848
			2	0.0020	0.9570	0.4049	−2.7578	0.0000	0.4024
		50	1	−0.0003	0.9590	0.3516	−1.9691	0.0000	0.3460
			2	−0.0000	0.9620	0.4957	−2.8181	0.0000	0.4888
Weighting	0.3	25	1	−0.0021	0.9310	0.2626	−0.0037	0.9370	0.2736
			2	0.0033	0.9400	0.3693	0.0007	0.9360	0.3835
		50	1	−0.0015	0.9340	0.3237	−0.0021	0.9390	0.3710
			2	0.0020	0.9370	0.4551	0.0014	0.9390	0.5184
	0.5	25	1	−0.0031	0.8820	0.2236	−0.0029	0.8950	0.2331
			2	0.0017	0.9060	0.3140	0.0004	0.9030	0.3266
		50	1	−0.0032	0.9070	0.2756	−0.0015	0.9180	0.3160
			2	−0.0004	0.9250	0.3868	0.0035	0.9080	0.4417

Table 2. Continued.

Method	Correlation	% mis	Y	MCAR bias	MCAR coverage	MCAR CI width	MARleft bias	MARleft coverage	MARleft CI width
Sample imputation	0.3	25	1	−0.0021	0.9650	0.2959	−0.0040	0.9520	0.3037
			2	0.0010	0.9460	0.4125	0.0076	0.9480	0.4293
		50	1	0.0004	0.9540	0.3422	−0.0062	0.9430	0.4281
			2	0.0005	0.9550	0.4879	0.0106	0.9540	0.6103
	0.5	25	1	−0.0016	0.9650	0.2824	−0.0014	0.9460	0.2905
			2	0.0013	0.9440	0.3954	0.0009	0.9450	0.4077
		50	1	0.0005	0.9540	0.3422	−0.0044	0.9540	0.3842
			2	0.0007	0.9550	0.4879	0.0098	0.9390	0.5402
Mass imputation	0.3	25	1	0.0003	0.9450	0.3857	−0.0200	0.9510	0.5419
			2	0.0005	0.9590	0.5480	0.0268	0.9460	0.7713
		50	1	−0.0008	0.9390	0.4772	−0.0237	0.9570	0.6636
			2	0.0030	0.9560	0.6752	0.0229	0.9440	0.9117
	0.5	25	1	−0.0001	0.9570	0.3289	−0.0051	0.9490	0.4663
			2	0.0007	0.9630	0.4603	0.0423	0.9400	0.6507
		50	1	−0.0033	0.9390	0.4033	−0.0051	0.9530	0.5743
			2	−0.0010	0.9470	0.5665	0.0438	0.9620	0.8202

Note that results of two target variables Y_1 and Y_2 are shown, which both have their own mean and variance, as illustrated in Subsection 3.1.

4.4. Overall Efficiency

We investigate efficiency of the methods in the sense that we investigate which methods have the smallest confidence interval widths under which conditions. When investigating the results, we see that CCA is an efficient method yielding valid inference under MCAR. There is no need for handling the nonresponse as the nonresponse is perfectly ignorable: the set of observed values can simply be analyzed to obtain unbiased estimates about the population. Even though the missingness is MCAR, treating the missingness can increase the statistical power of the analyses at hand. This is demonstrated by weighting and imputing the sample as the confidence intervals under these approaches are generally more narrow than under CCA. Mass imputation, on the other hand, does not show this result. This can simply be explained by the severity of the problem that is considered with mass imputation in our simulation setup. After all, under mass imputation we aim to solve at least a 96.25% missingness problem.

Even though mass imputation may yield less sharp inference than sample imputation and weighting, the inference is valid and exhibits correct variance properties under all simulation conditions. The same can be said of sample imputation, but with much sharper inference. The estimates obtained under weighting are unbiased, the intervals are among the smallest, but the coverage rates are somewhat low. Especially when larger correlations occur in the data, one could question the validity of inference obtained by weighting. Furthermore, it is surprising that these low coverage rates occur under both MCAR and MAR, indicating that the variance of a weighted mean estimated using Taylor linearization indeed ignores uncertainty about the missing data and possibly about the weights as well.

5. Discussion

We have demonstrated that weighting and imputation are practically equivalent when unbiased estimation is of interest. However, the inference obtained under weighting may be questionable in situations where multiple imputation approaches exhibit correct variance properties and well-covered population estimates. In general it holds that inferring about the population by imputing the sample yields efficient, unbiased estimates in all simulated conditions, which is in line with conclusions drawn by Peytchev (2012).

A main characteristic of our simulation approach is that it deals with a SRS_{WOR}. With more complex sampling approaches, it would not be sufficient to only impute the sample, since the complex sampling structure is then ignored. Although we did not investigate this, we do expect that mass imputation will lead to unbiased and efficient estimates when a more complex sample is drawn because the design of the complex sample is always based on observed information, so the missingness mechanism describing the sample to the population is always MAR. However, this is not included in this simulation study, and additional research should be done.

Furthermore, in this simulation we assume quite an ideal situation, where the sample is perfectly linked to a completely observed population register. Of course, this is not often the case in practice. In addition to the traditional Total Survey Error framework introduced

by Groves et al. (2009), Zhang (2012) introduced a two-phase life cycle of integrated statistical micro data, which also discusses the errors that might be encountered when multiple data sets are combined, such as identification or comparability error. Furthermore, we also assume that our population register is perfectly observed. This is in practice also not often the case, although this is commonly assumed by many researchers. Recently, imputation methods have been developed to take misclassification in combined data sets into account, for example by assuming that a certain proportion of the data is misclassified (Manrique-Vallier and Reiter 2016) or by estimating the number of misclassified units by using information from multiple sources (Boeschoten et al. 2016).

It is clear that weighting does not include all sources of uncertainty. This limits the validity of the inference obtained under weighting. Theoretically, these sources of uncertainty could be added to the estimations that are obtained from weighted data sets. However, we have demonstrated that the imputation approaches take the sources of variations about the observed and missing data properly into account. Adjusting the weighted estimation to allow for valid inference under unit nonresponse would therefore be redundant as it is a complicated step to solve a problem that can be straightforwardly solved by another approach.

In addition, weighting cannot handle partial response (Van Buuren 2012, 22). Analyzing multivariate response data with partial responses will be particularly problematic when weighting is applied, and multiple imputation is a very suitable alternative in this setting.

It is known that complete case analysis yields valid inference under MCAR mechanisms and that its performance may be severely impaired under MAR missingness. The results of complete case analysis in simulations can be very informative, as it can act as a point of reference for the performance of other methods. At the same time, the validity of the simulation scheme can be assessed, because we know the theoretical properties under which complete case analysis can be applied. Failure to meet these expectations indicates a faulty simulation scheme. This is not the case.

The simulation study conducted in this article illustrated that multiple imputation methods lead to valid inference in situations of unit nonresponse and have practical advantages over weighting. Whether sample or mass imputation methods should be used depends on the specific data structure.

6. References

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys,* volume 568 of *Wiley Handbooks in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Boeschoten, L., D. Oberski, and T. de Waal. 2016. "Estimating Classification Error under Edit Restrictions in Combining Survey-Register Data." *Journal of Official Statistics* 33:921–962. Doi: http://dx.doi.org/10.1515/JOS-2017-0044.

Cohn, N. 2016. "How One 19-Year-Old Illinois Man is Distorting National Polling Averages." *The New York Times*. Available at: https://nyti.ms/2k5sB5z (accessed September 26, 2017).

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation,* volume 563 of *Wiley Handbooks in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American statistical Association* 87: 376–382.

Groves, R.M., F.J. Fowler, Jr, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology,* volume 561 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Howard, W.J. 2012. *Using Principal Component Analysis (pca) to Obtain Auxiliary Variables for Missing Data in Large Data Sets*. University of Kansas. PhD Dissertation.

Lumley, T. 2014. *Analysis of Complex Survey Samples*. Available at: http://cran.r-project.org/web/packages/survey/survey.pdf (accessed September 26, 2017).

Manrique-Vallier, D. and J.P. Reiter. 2016. "Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data." *Journal of the American Statistical Association*. Doi: http://dx.doi.org/10.1080/01621459.2016.1231612.

Peytchev, A. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly* 76: 214–237. Doi: https://doi.org/10.1093/poq/nfr065.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, USA.

Rubin, D.B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91: 473–489.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag.

Schulte Nordholt, E., J. van Zeijl, and L. Hoeksma. 2014. *Dutch Census 2011, Analysis and Methodology*. The Hague/Heerlen. Available at: https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf (accessed 26 September 2017).

Stapleton, L.M. 2008. "Analysis of Data from Complex Surveys." In *International Handbook of Survey Methodology*, edited by E.D. De Leeuw, J.J. Hox, and D. Dillman, 342–369. Psychology Press, Taylor & Francis Group, New York.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. CRC press, Boca Raton, Florida.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45: 1–67. Doi: http://dx.doi.org/10.18637/jss.v045.i03.

Vink, G. and S. van Buuren. 2014. "Pooling Multiple Imputations when the Sample Happens to be the Population." *arXiv preprint arXiv:1409.8542*. Available at: https://arxiv.org/pdf/1409.8542.pdf (accessed 26 September 2017).

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016. "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation." *Biometrics* 72: 242–252.

Received November 2015
Revised January 2017
Accepted March 2017

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 979–1004, http://dx.doi.org/10.1515/JOS-2017-0046

The Effects of the Frequency and Implementation Lag of Basket Updates on the Canadian CPI

Ning Huang[1], *Waruna Wimalaratne*[1], *and Brent Pollard*[1]

In this article, we examine the effects of different frequencies and implementation months of basket updates on the fixed-basket price index – the Lowe index, through theoretical analysis and empirical simulation using Canadian data from 2000 to 2013. We find that both an increased frequency of basket updates and a faster implementation of these new baskets will reduce substitution bias in the CPI. However, we also find that improvements to the method of accelerating frequency has diminishing marginal returns in practice – as each subsequent increase in the frequency with which the CPI basket is updated has a less pronounced effect; and the ideal link-month when a new basket is implemented is unpredictable, since the impact of the implementation lag depends upon the consistency between short-term price movements and long-term price trends.

Key words: Consumer price index; Lowe formula; fixed-basket index; basket-update frequency; implementation lag; measurement bias; commodity substitution bias; superlative indexes; Fisher, Walsh, Törnqvist indexes.

1. Introduction

The Consumer Price Index (CPI) is the most widely used indicator of price change in Canada. It serves a variety of purposes, and is therefore of interest to governments, unions, business organizations, research institutions, and the general public. Its various uses include its function as a general indicator of inflation in Canada and as a tool for adjusting incomes, wages, and other payments to ensure that purchasing power is unaffected by any average price movement. Further details are available in a reference paper published by Statistics Canada (2014).

In line with the practices of most other national statistical agencies, the Consumer Prices Division (CPD) at Statistics Canada uses the Lowe index formula for aggregating its CPI at the upper level. The Lowe index formula, often described as a "Laspeyres-type" formula, is a fixed-basket formula. This means that the quantity and quality of the goods

[1] Statistics Canada, Consumer Prices Division, 170 Tunney's Pasture, Ottawa, O.N., K1A 0T6. Canada. Emails: Nings.Huang@Canada.ca, Waruna.Wimalaratne@Canada.ca, and Brent.Pollard@Canada.ca.

Acknowledgments: The authors thank Ross Beck-MacNeil, Nathalie Brault, Kyle de March, Xin Ha, Mathieu Lequain, Sue Morris, Marc Prud'homme, Bradley Snider, Philip Smith, Faouzi Tarkhani, Amanda Wright, Alice Xu, Clément Yélou, two internal reviewers at Statistics Canada, and Price Measurement Advisory Committee members Jan de Haan, Erwin Diewert, Pierre Duguay, David Fenwick, John Mallon, Marshal Reinsdorf, Patrick Sabourn, Mick Silver, Kam Yu for their helpful discussions and comments. Thanks also to three anonymous reviewers for their comments. The views expressed in this article are solely those of the authors and do not necessarily reflect those of Statistics Canada.

and services included in the CPI basket must be unchanged or equivalent within the life span of a CPI basket. It is also referred to as a Cost of Goods Index (COGI).

The Lowe formula is used in practice because it offers a simple and convenient way to compile composite price indexes in a timely manner. Although the Lowe formula is a good choice for the fixed-basket concept of a CPI, its inherent limitations must be taken into consideration; for example, it cannot account for consumer's price-induced product substitution, it experiences delay in reflecting the effects of new goods and services on consumer price change, and it has difficulty in fully accounting for changes in the quality of existing consumer products. Due to these and other limitations, the official CPI, published by Statistics Canada, is not a true measure of actual changes in the cost of living.

A Cost-of-Living Index (COLI) is derived from the standpoint of an economic theory, based upon the assumption of a household's utility optimization behaviour, which assumes that a household will structure its purchases to maximize utility, or satisfaction, given a certain level of prices and a certain level of income. Since a household's utility optimization problem is dual to its cost minimization problem, a COLI then measures the change in the household's minimum cost of maintaining a fixed level of utility over two periods when faced with changes in prices. The theory of the COLI provides the conceptual framework for some countries' CPI, such as the United States (U.S.) and Sweden.

The difference between the official CPI and an underlying COLI, which can be approximated by a class of superlative indexes, is called measurement bias. According to the Consumer Price Index (CPI) Manual (ILO et al. 2004), a group of "superlative" price indexes, such as the Fisher, Walsh and Törnqvist indexes, is expected to provide "fairly close" approximations to the underlying COLI. Thus, they are recommended in the manual as the "target indexes" for the upper-level index. The main types of measurement bias include commodity-substitution bias, outlet-substitution bias, quality-change bias and new-goods bias. In this study, only the measurement bias associated with the upper-level aggregation is discussed. Apart from this bias, there could also be sampling and other non-sampling bias in the estimated elementary indexes and estimated basket weights. Note that measurement bias can be measured in terms of index level and index growth rate. In this article, commodity-substitution bias is analyzed and reported in both ways depending on the context of the article.

These measurement biases arise from the fact that any basket weights, held constant over more than one period, do not necessarily reflect the types of purchases that consumers actually make to attain the same level of welfare when relative prices change. A fixed-basket index, therefore, normally fails to account for the changes in consumers' purchasing patterns or preferences in a timely manner, and measures only the average price movement based on a specifically defined basket, resulting in measurement bias.

A COLI, on the other hand, allows for changes in the basket over time and, therefore, accounts for changes in consumer purchasing patterns when measuring average price movements over two periods. While numerous national statistical offices do not construct their CPIs as a COLI, including Statistics Canada, many of them still want to have knowledge about the measurement bias in their official CPI because of its important role as a major economic indicator and as a wage or salary indexation factor.

Since the CPI is the most commonly used indicator for tracking overall price change in Canada, measurement bias in the CPI is an important issue for both its users and compilers.

As Sabourin (2012) pointed out, "since the CPI departs from a true COLI, it is subject to measurement bias and does not necessarily reflect changes in the wellbeing of consumers, which could be problematic for monetary policy and when making cost-of-living adjustments to wages and salaries."

Given the varying uses of the CPI, research on the measurement bias in the Canadian CPI is conducted regularly by some of its users, such as continuous research conducted by the Bank of Canada, including Crawford (1998), Rossiter (2005) and Sabourin (2012). According to Sabourin (2012), for the years from 2005 to 2011, the mean total bias in the Canadian CPI was 0.45 percentage points per year from 2005 to 2011, among which commodity-substitution bias was 0.22, outlet substitution bias was 0.04, new-good bias was 0.20 and quality adjustment bias was −0.01. Similar studies quantifying the bias in the CPI have been conducted in other countries, such as the paper by Boskin et al. (1996), also known as the Boskin Commission Report, for the U.S., which stated that "the Commission's best estimate of the size of the upward bias looking forward is 1.1 percentage points per year. The range of plausible value is 0.8 to 1.6 percentage points per year." The estimates of CPI bias can also be found in Shiratsuka (2006) for Japan and in Wynne and Rodriguez-Palenzuela (2002) for European countries.

This article focuses on the investigation of commodity-substitution bias, which is caused by the inability of a fixed-basket index to capture consumers' price-induced substitution. Generally speaking, without changing the formula for compiling the CPI, this type of bias could be reduced by updating the CPI basket more frequently and by implementing the basket in a more timely fashion. Both of these methods allow a more accurate reflection of the changes in purchasing patterns due to consumers' substitution between different combinations of goods and services. In the existing literature associated with commodity-substitution bias, there are only a limited number of studies examining the impact on the CPI of the frequency and delay of implementing new basket weights. This is likely due to the difficulties associated with acquiring such data. In the Canadian context, the annual household expenditure survey facilitated this study.

It is widely recognized that more frequent basket updates and faster implementation will lead to an index that more closely approximates a superlative measure. For instance, Japan publishes two series of CPI: the official CPI, with weights updated every five years; and a chained Laspeyres CPI, with weights updated annually.

A study by Greenlees and Williams (2009) showed that quarterly weight updates generated an index that more closely resembled a target index when compared to less frequent updates. In their study, a chained Törnqvist index was calculated as a superlative target. They simulated various weight updating periods: quarterly, semi-annual, annual and biennial. The index derived from quarterly weights approximated most closely to the superlative index. They also found that the Lowe index updated annually, which could be realistically compiled under the operational constraints, increased less than the rolling, two-year index of current methodology in four out of the six years studied (2002 to 2007). In addition, the advantage of using more timely weights was not offset by any increase in index volatility or instability.

Ho et al. (2011) examined, using data from 2002 to 2008, the impact on the New Zealand CPI of reweighting at different frequencies and at different levels of the index structure. They showed that frequent weight updates at the sub-item level and above generated CPI

series that tracked the Fisher series most closely among those generated by using other weight-update frequencies and other aggregation levels. Their current methodology with weight updates in June 2002, 2006, and 2008 quarters yielded a Laspeyres index of 117.0, while their methodology without updates produced an index of 117.9; these can be compared to a Fisher of 115.8, for the June 2008 quarter.

In addition to the frequency of basket updates, national statistical offices also need to determine when to introduce a new basket. The delay in the implementation of a new basket affects the size of commodity-substitution bias. Limited research supports this: Généreux (1983), using Canadian data, compared a chained Laspeyres series with eight basket updates against a chained Laspeyres series with only one basket update over the period from 1957 to 1978. He concluded "what appears to be desirable is not necessarily a more frequent updating of the CPI baskets but a more timely one." For example, implementing the new weights in the years they refer to could considerably reduce the commodity-substitution bias. Using Canadian data, Bérubé (1996) also showed that introducing a basket two years after the basket reference period would reduce the annual substitution bias from 0.20 percentage points to 0.18 percentage points over the period from 1962 to 1994, compared with introducing a basket three years after the reference period.

A study from Australia Bureau Statistics (ABS 2016) showed a significant decrease in substitution bias by having shorter weight implementation lag for the period between September 2005 and September 2011. The bias declined from 0.24% per year for the CPI to 0.09%, 0.15%, and 0.16% with weight implementation lags of one, two, and three years, respectively. The Australian CPI weights are updated every six years using a household survey. In their study they utilized household final consumption expenditure from National Accounts to calculate the Lowe Index.

In 2010, Statistics Canada implemented the *CPI Enhancement Initiative*, a multi-stage program to advance the quality of the CPI. As part of this initiative, effort was directed at identifying and reducing the commodity-substitution bias. In 2013, a more frequent basket update schedule was implemented – from once every four years to once every two years. Additionally, the 2011 basket was introduced more quickly than past baskets – the time lag went from 16 months to 13 months. Interest and focus subsequently shifted to investigating the effect that changes such as these have on the quality of the CPI. The results would help inform the decision of whether to further accelerate the frequency of basket updates and further reduce the implementation lag.

The Canadian economy, similar to those of other major economies, is a knowledge-based economy, associated with dynamic technological change. With the rapid applications of new technology and emergence of new products and new market structure, consumers' lifestyles and merchants' pricing strategies have also experienced significant change. As a result, it is expected that a CPI basket becomes outdated more rapidly.

This in turn, raises questions for compilers of CPIs attempting to improve index quality and accuracy: does the comparison by Généreux (1983) between "a more frequent updating of the CPI baskets" and "a more timely one" still hold? Are empirical results from other countries, such as those revealed by Greenlees and Williams (2009) also valid for Canada? And, how can national statistical offices reduce the commodity substitution bias further?

Updating the basket weights of a price index such as the CPI is accomplished in various stages. How each of these are implemented will likely have some effect on the overall

index. This article will focus on the performance of the index under different scenarios for two of these stages, the weight-updating cycle, and the timeliness of the introduction of the new weights. The principal source of the data for the study is the Canadian Survey of Household Spending, which is used to reflect changes in consumers' spending patterns over time, for the period from 2002 to 2013. Price indexes from the Canadian CPI are also used.

To estimate the substitution bias, this article compares the results of the Lowe price index with those of the Fisher price index. This approach differs from the more common method of estimation, which compares the results between the Laspeyres price index and the Fisher price index. Another difference lies in the focus of the analysis: instead of only reporting the empirical results derived from Canadian data, the divergence in the resulting indexes obtained under various scenarios (different weight-updating schedules, and different implementation lags of the introduction of new weights) is analyzed in detail using a mathematical approach. Consequently, this article will shed new light on how to mitigate the well-known and pervasive substitution bias which characterizes a fixed-basket CPI for national statistical offices in countries facing similar situations as Canada. For these countries, the findings will play an important role in determining the desired frequency of weight updates and time of implementing new weights.

The remainder of this article is organized as follows: Section 2 discusses the data sources and data construction methods; Section 3 defines the target price index, which belongs to a group of superlative series that closely approximate a COLI, used in this study; Section 4 addresses the effects of the frequency and implementation lag of weight updates on the Canadian CPI in detail; and Section 5 concludes the article.

2. Data Construction

The two main elements required for the calculation of a price index series are prices and quantities. To this end, this study makes use of two main sources of data – the Consumer Price Index (CPI) and the Survey of Household Spending (SHS). The CPI provides data on the price indexes for each of its measured goods and services at the basic class level of aggregation. Basic classes are the lowest-level aggregates of products, chosen by Statistics Canada, for which a set of weights is fixed for the duration of the CPI basket. The SHS data are used in constructing fixed-basket weights for twelve years going from 2000 to 2011 based on the 2005 CPI classification structure. In this way, the estimated substitution bias would not be affected by the impact of changes in the specification and the appearance of new products.

The "price" component of the index calculation comes from the CPI over the period from January 2000 to December 2013. The original price indexes are unlinked price indexes for each of the corresponding published CPI basket. To facilitate the index reconstruction, the indexes were linked together based on the classification of the 2005 basket and rebased to January 2000 = 100. The reconstructed indexes, therefore, represent the price movement from the price reference period of January 2000 to a given price observation month.

The "quantity" component of the index comes from the SHS, which contains detailed information about consumer spending during a given reference year. The SHS sample has

a cross-sectional design, and is selected from the Labour Force Survey sampling frame and carried out in private households. The SHS is the main source of the expenditure weights data for the CPI.

In the first stage of the data construction, we derived expenditure weights for the years without official CPI weights – 2000, 2002 to 2004, 2006 to 2008, and 2010, using data from the SHS. Official CPI weights data were used whenever they were available; specifically the 2001, 2005, 2009, and 2011 baskets. However, some adjustments were made in order to align them with the 2005 classification of the CPI at the basic class level of aggregation. The 2005 classification structure that was in use in the official CPI from May 2007 to April 2011 was maintained across time to preserve uniformity and avoid complications arising from the introduction of new items. For non-official basket update years, some expenditure values were unavailable from the SHS; for example, the low level details for some basic classes under the food classification. To estimate these expenditure values, we used a modified price-updating method which used a weighted average of expenditures for those years with detailed SHS information. With this method, relatively greater importance is assigned to expenditures in baskets from periods closer to the imputed period. For example, the unknown expenditure for item i in 2003 can be imputed from the formula:

$$p_i^{2003}q_i^{2003} \equiv (6/8)\underbrace{p_i^{2001}q_i^{2001}\frac{p_i^{2003}}{p_i^{2001}}}_{\text{upward price update}} + (2/8)\underbrace{p_i^{2009}q_i^{2009}\frac{p_i^{2003}}{p_i^{2009}}}_{\text{backward price update}} \quad (1)$$

Finally, similar imputation strategies were employed for calculating the weights for the mortgage interest cost basic class as well as some components of the clothing classification. In the case of the mortgage interest cost index, where Statistics Canada has a special treatment, data were available only for the official basket reference years. As a remedy, weights for the remaining years were calculated using the same method as employed for food classification. For the replacement cost basic class, the SHS lacked detailed housing data for non-official basket update years, and so a combination of internal and external data was used to calculate its value.

Once the "price" and "quantity" components were built, a data validation was performed by reconstructing the official CPI using the analytical database. Comparing the constructed CPI with the official CPI, we believe that the analytical series was a very good approximation.

3. Target Index Formula

To determine the magnitude of the commodity-substitution bias, first, it is necessary to select a target index with which to compare the estimates of this study. The Fisher, Walsh, and Törnqvist indexes have been widely used for this purpose, as they belong to a small class of "superlative indexes".

An important characteristic of superlative indexes is that they include the prices and quantities in both periods being compared, they are therefore symmetrically weighted indexes. Moreover, these three index number formulas are flexible and provide second-order approximation to each other. In other words, different superlative indexes tend to

have similar properties, yield similar results and behave in very similar ways. In addition, they are expected to provide a close approximation to the underlying conditional cost-of-living index (COLI). Diewert (1976) showed that superlative indexes provide close approximations to any true cost-of-living price index if the underlying utility function is linear homogeneous. As a close approximation to the unknown COLI, superlative indexes are recommended in the CPI ILO Manual as the theoretical target indexes. The difference between the Laspeyres-type index, which does not permit the commodity-substitution induced by relative price changes, and the target indexes can be treated as a measure of commodity-substitution bias at the upper level of index aggregation when holding classification structure unchanged.

In this study, we aim at comparing chained-CPI series constructed by applying different weights. The target indexes are, therefore, estimated by using the chain-linked Fisher, Walsh, and Törnqvist index number formulas with annual weight-updating, as detailed monthly expenditure data are unavailable. The corresponding annual CPI series are derived by taking the unweighted arithmetic average of monthly price indexes of the twelve months in the calendar year. Using the Fisher index number formula $P_{\mathrm{ChF}}^{(2003+t)/2003}$ as an example, we show how the chain-linked index between 2003 and 2011 is constructed:

$$
\begin{aligned}
P_{\mathrm{ChF}}^{(2003+t)/2003} &= \prod_{j=1}^{t} P_{\mathrm{F}}^{(2003+j)/(2003+j-1)} \\
&= \prod_{j=1}^{t} \left(\underbrace{\frac{\sum_{i=1}^{N} p_i^{2003+j} q_i^{2003+j-1}}{\sum_{i=1}^{N} p_i^{2003+j-1} q_i^{2003+j-1}}}_{P_{\mathrm{L}}^{2003+j/2003+j-1}} \underbrace{\frac{\sum_{i=1}^{N} p_i^{2003+j} q_i^{2003+j}}{\sum_{i=1}^{N} p_i^{2003+j-1} q_i^{2003+j}}}_{P_{\mathrm{P}}^{2003+j/2003+j-1}} \right)^{1/2} \qquad (2) \\
& \quad t = 1, 2, \ldots 8
\end{aligned}
$$

where $P_{\mathrm{ChF}}^{(2003+t)/2003}$ denotes chained-Fisher from 2003 to 2003 + t; $P_{\mathrm{F}}^{(2003+j)/(2003+j-1)}$ denotes the direct Fisher index from 2003 + j − 1 to 2003 + j; $P_{\mathrm{L}}^{(2003+j)/(2003+j-1)}$ denotes the direct Laspeyres index from 2003 + j − 1 to 2003 + j and $P_{\mathrm{P}}^{(2003+j)/(2003+j-1)}$ denotes the direct Paasche index from 2003 + j − 1 to 2003 + j. N is the total number of goods and services included in the CPI basket. The chained-Walsh index and chained-Törnqvist index can be compiled similarly. The three superlative indexes are expected to behave similarly, which is confirmed by the numerical results over the period from 2003 to 2011 reported in Table 1, where the average growth rate using the chained Fisher index as an example, is calculated as $\sqrt[8]{\left(P_{\mathrm{ChF}}^{2011/2003}/100\right)} - 1$.

In the next section, the target index values in Table 1 can be compared with the CPI series compiled with different CPI weight-updating schedules to produce the estimates of the upper-level commodity-substitution bias. More specifically, the chained-Fisher index is used as an example to estimate the commodity-substitution bias in this article.

Table 1. Superlative price indexes (2003 = 100).

Year (2003 + t)	Fisher		Walsh		Törnqvist	
	Chained index	Annual inflation	Chained index	Annual inflation	Chained index	Annual inflation
2003	100.000		100.000		100.000	
2004	101.728	1.728	101.730	1.730	101.730	1.730
2005	103.746	1.984	103.750	1.986	103.750	1.986
2006	105.475	1.667	105.480	1.668	105.482	1.669
2007	107.401	1.826	107.409	1.829	107.410	1.828
2008	109.624	2.069	109.632	2.070	109.633	2.069
2009	109.670	0.042	109.684	0.047	109.688	0.050
2010	111.404	1.581	111.422	1.585	111.422	1.581
2011	114.389	2.679	114.408	2.680	114.405	2.677
Average growth rate (2003–2011)		1.695		1.697		1.696

4. Approaches to Reducing Commodity-Substitution Bias

In general, the commodity-substitution bias could be measured as the difference between the published CPI and the target index, both of which are estimated by keeping the items in the baskets fixed over time. The source of this substitution bias varies. Two important sources could be the frequency of CPI basket weight updates, and the time lag between the end of the basket reference year and the initial implementation time of a new CPI basket in the CPI calculation. Using the 2011 basket update as an example, we illustrate the relationship among different time periods involved in the index calculation using the following timeline.

On the timeline in Figure 1, the basket reference year (during which the SHS is conducted to collect the necessary information for the CPI basket) is 2011. The 2011 CPI basket was implemented with the February 2013 CPI, which is defined as the implementation month in this paper. The duration from January 2012 to January 2013 is the implementation lag, which in this case is 13 months. January 2013 is the link month for the implementation of the 2011 CPI basket.

In this section, how the frequency and implementation lag of the CPI weight affect the magnitude of the upper-level commodity-substitution bias will be explored in further detail.

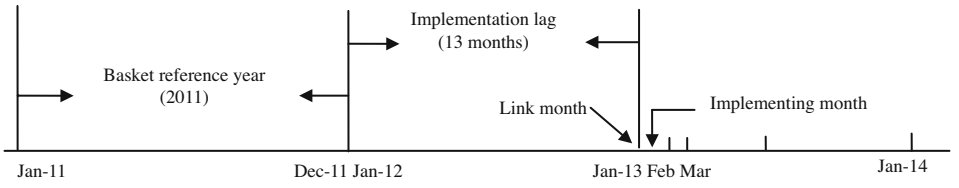

Fig. 1. Timeline of CPI basket update.

4.1. Commodity-Substitution Bias and the Frequency of Basket Updates

4.1.1. Conceptual Framework to Measure the Impact of the Basket Update Frequency

The CPI basket is designed to reflect consumers' spending patterns. As a result of both relative price changes and some long-term effects on consumers' spending behaviour, such as the impact of demographic factors and technological changes, the weights might become out-of-date and less representative of current consumption patterns. The bias in a Lowe index is likely to increase as the basket weights age. Therefore, CPI weights should be updated periodically to reflect the changes in these patterns.

To identify the pure impact of the frequency of weight updates on the magnitude of the CPI bias, we fix the implementation lag at 13 months and vary only the frequency of weight updates when calculating the All-items CPI, which measures price change of all the goods and services included in the Canadian CPI, for the period from January 2002 to December 2013.

A direct Lowe index $P_{\mathrm{Lo}}(p^0, p^t, q^b)$ can be defined in terms of a quantity vector $q^b \equiv \left[q_1^b, \ldots, q_N^b\right]$, a price vector of base period $p^0 \equiv \left[p_1^0, \ldots, p_N^0\right]$ and a price vector of current period $p^t \equiv \left[p_1^t, \ldots, p_N^t\right]$:

$$P_{\mathrm{Lo}}(p^0, p^t, q^b) = \frac{\sum_{i=1}^{N} p_i^t q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} \tag{3}$$

where N is the total number of goods and services included in the CPI weight structure.

It can be also written in terms of the hybrid share form as follows:

$$\begin{aligned} P_{\mathrm{Lo}}(p^0, p^t, s^{0:b}) &= \frac{\sum_{i=1}^{N} p_i^t q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} = \sum_{i=1}^{N} \left(\frac{p_i^t}{p_i^0}\right) \frac{p_i^0 q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} \\ &= \sum_{i=1}^{N} \left(\frac{p_i^t}{p_i^0}\right) s_i^{0:b} \end{aligned} \tag{4}$$

where the hybrid expenditure shares $s_i^{0:b}$ corresponding to the quantity weights vector q^b measured at base period price vector p^0 are defined as:

$$s_i^{0:b} = \frac{p_i^0 q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b}, \quad i = 1, 2, \ldots, N \tag{5}$$

If more than one basket, say baskets $b1$ and $b2$, are in use, it is necessary to calculate the chain-linked Lowe index, where the indexes calculated using different CPI baskets are linked together. To explain this concept, let $p^{y,m}$ be the elementary price vector for year $y \geq 2002$ and month $m = 1, 2, \ldots, 12$; the chain-linked Lowe index for year y and month m, with every x years as the frequency of weight updates, is denoted as $P_{\mathrm{ChLo}_x}(y, m)$. The calculation of the chain-linked Lowe index depends on which basket is currently used and in which month it is linked to the previous basket. In general,

a chain-linked Lowe index can be defined as:

$$P_{\mathrm{ChLo}_x}(y,m) = P_{\mathrm{ChLo}_x}(link\_month)\, P_{\mathrm{Lo}}(p^{link\_month}, p^{y,m}, q^{b}) \tag{6}$$

where $P_{\mathrm{ChLo}_x}(link\_month)$ is a chain-linked Lowe index for the link month that chains together indexes using the current basket q^{b} and the previous baskets.

If the CPI basket is assumed to be updated every x years, where x can be 1, 2, 3, 4, or 5, after the adoption of the 2000 basket, the Equation (6) can be applied to compile the CPI series. With the implementation lag set equal to 13 months, a new basket $2000 + kx$ is introduced in February of year $2002 + kx$ with January ($m = 1$) of year $2002 + kx$ as the link month, $k = 1, 2, \ldots$ such that $2002 + kx \leq y$ (y is the year of the price index). With these assumptions, the chain-linked Lowe index can be calculated by substituting the corresponding values in Equation (6), yielding the following results:

$$P_{\mathrm{ChLo}_x}(y,m) = P_{\mathrm{ChLo}_x}(2002 + kx, 1)\, P_{\mathrm{Lo}}(p^{2002+kx,1}, p^{y,m}, q^{2000+kx}) \tag{7}$$

The first component of the right-hand side of the Equation (7), $P_{\mathrm{ChLo}_x}(2002 + kx, 1)$, is the link factor, which is also a chain-linked Lowe index for January of year $2002 + kx$, which is the link month of the current basket ($2000 + kx$); the second component, $P_{\mathrm{Lo}}(p^{2002+kx,1}, p^{y,m}, q^{2000+kx})$, is the direct Lowe index comparing the current month (y, m) with the link month ($2002 + kx$, *1*), January of year $2002 + kx$.

The link factor $P_{\mathrm{ChLo}_x}(2002 + kx, 1)$ can be also defined as the product of several direct Lowe indexes as follows:

$$\begin{aligned} P_{\mathrm{ChLo}_x}(2002 + kx, 1) = {} & P_{\mathrm{Lo}}(p^{0}, p^{2002+x,1}, q^{2000}) \\ & P_{\mathrm{Lo}}(p^{2002+x,1}, p^{2002+2x,1}, q^{2000+x}) \\ & \cdots P_{\mathrm{Lo}}(p^{2002+(k-1)x,1}, p^{2002+kx,1}, q^{2000+(k-1)x}) \end{aligned} \tag{8}$$

where k denotes the number of times the CPI basket is updated since the price reference period, which is assumed to be during the life span of the basket q^{2000}.

We now describe how the chain-linked Lowe index can be constructed if the weights are updated every two years, that is $x = 2$. Denote the chain-linked Lowe index for year y and month m, with an update frequency of every two years, by $P_{\mathrm{ChLo}_2}(y,m)$. In this case, the direct Lowe index, which uses the 2000 basket only, is employed from February 2002 to January 2004, with January 2002 as the link month, the overlapping period that links the old and new CPI series. Applying (9), we have $P_{\mathrm{ChLo}_2}(2002, 1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2002,1}, q^{2000}) = 1$. Thus, the chain-linked Lowe index defined by (9) is, for the first 24 months running from February 2002 to January 2004, equal to the direct Lowe index:

$$\begin{aligned} P_{\mathrm{ChLo}_2}(y,m) = {} & P_{\mathrm{Lo}}(p^{2002,1}, p^{y,m}, q^{2000}) \\ & (\text{with } y = 2002,\ 2003;\ \text{and } m = 1, 2, \ldots, 12 \text{ and } y = 2004;\ m = 1) \end{aligned} \tag{9}$$

The same direct Lowe index on the right hand side of (9) is, therefore, used to define the chain-linked Lowe index for January 2004:

$$P_{\mathrm{ChLo}_2}(2004,1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000}) \tag{10}$$

The above chain-linked Lowe index for January 2004 corresponds to the link factor that chains together indexes using the 2000 basket and the 2002 basket. For the remaining months in 2004 and 2005, the annual quantity weights vector q^{2002} becomes available and the chain-linked Lowe index is defined as follows:

$$\begin{gathered} P_{\mathrm{ChLo}_2}(y,m) = P_{\mathrm{ChLo}_2}(2004,1)\, P_{\mathrm{Lo}}(p^{2004,1}, p^{y,m}, q^{2002}) \\ (\text{with } y = 2004,\ 2005;\ m = 1,2,\ldots,12;\ y = 2006;\ m = 1) \end{gathered} \tag{11}$$

The chain-linked Lowe index for January 2006 is, therefore, defined as follows:

$$P_{\mathrm{ChLo}_2}(2006,1) = P_{\mathrm{ChLo}_2}(2004,1)\, P_{\mathrm{Lo}}(p^{2004,1}, p^{2006,1}, q^{2002}) \tag{12}$$

Here again, the chain-linked Lowe index for January 2006 is the link factor that chains indexes based on 2004, 2002, and 2000 baskets respectively. From February 2006 to January 2008, the annual quantity weights vector q^{2004} becomes available and the chain-linked Lowe for this time span is defined as follows:

$$\begin{gathered} P_{\mathrm{ChLo}_2}(y,m) = P_{\mathrm{ChLo}_2}(2006,1)\, P_{\mathrm{Lo}}(p^{2006,1}, p^{y,m}, q^{2004}) \\ (\text{with } y = 2006,\ 2007;\ m = 1,2,\ldots,12;\ y = 2008;\ m = 1) \end{gathered} \tag{13}$$

Once more, the link factor chaining the indexes together across baskets is the chain-linked Lowe index for January 2008 which continues to be defined by the right-hand side of (13), as follows:

$$P_{\mathrm{ChLo}_2}(2008,1) = P_{\mathrm{ChLo}_2}(2006,1)\, P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004}) \tag{14}$$

Continuing the above process, we can construct the chain-linked Lowe index for other months in the other years.

To show how the defined process works, here we compile a chain-linked Lowe index for a particular month, say August 2011, as an example. Assume the weight-updating frequency is two ($x = 2$) and implementation lag is 13 months. The chained Lowe index is then denoted by $P_{\mathrm{ChLo}_2}(2011,8)$. Based on the described process, the current period, August 2011, is identified to be in the time span going from February 2010 to January 2012 and the associated quantity weights vector is q^{2008}, with January 2010 as the link month. The chain-linked Lowe index $P_{\mathrm{ChLo}_2}(2011,8)$ can then be constructed as:

$$P_{\mathrm{ChLo}_2}(2011,8) = P_{\mathrm{ChLo}_2}(2010,1)\, P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008}) \tag{15}$$

where $P_{\mathrm{ChLo}_2}(2010,1)$ is the link factor that chains together the price indexes using the 2008 basket and the previous baskets. Based on Equation (8), it can be written as a product

of direct Lowe indexes as follows:

$$P_{\mathrm{ChLo}_2}(2010,1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000})\, P_{\mathrm{Lo}}(p^{2004,1}, p^{2006,1}, q^{2002}) \\ P_{\mathrm{Lo}}(p^{2006,01}, p^{2008,1}, q^{2004})\, P_{\mathrm{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006}) \tag{16}$$

The direct Lowe index on the right-hand side of (15) can be compiled based on Equation (3) as follows:

$$P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008}) = \frac{\sum_i p_i^{2011,8}\, q_i^{2008}}{\sum_i p_i^{2010,1}\, q_i^{2008}} \tag{17}$$

Next, the chain-linked Lowe index for the same month, August 2011, but with different weight-updating frequency, $x = 3$, denoted by $P_{\mathrm{ChLo}_3}(2011,8)$, is considered. It can be compiled based on the process described in the case of a weight update every two years (refer to Equation (9) to (14)), as follows:

$$P_{\mathrm{ChLo}_3}(2011,8) = P_{\mathrm{ChLo}_3}(2011,1)\, P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009}) \tag{18}$$

With the two CPI index values associated with different frequencies of weight updates, the commodity-substitution bias can be then estimated by comparing the chain-linked Lowe index with the same target index. For example, let $Bias_{\mathrm{ChLo}_2}(2011,8)$ and $Bias_{\mathrm{ChLo}_3}(2011,8)$ denote the commodity-substitution bias, measured in terms of index level, of the chain-linked Lowe index for August 2011, with weight-updating frequencies equal to every two and every three years, respectively. They can be defined as follows

$$Bias_{\mathrm{ChLo}_2}(2011,8) = P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{Target}}(2011,8) \tag{19}$$

$$Bias_{\mathrm{ChLo}_3}(2011,8) = P_{\mathrm{ChLo}_3}(2011,8) - P_{\mathrm{Target}}(2011,8) \tag{20}$$

To compare the magnitude of the bias generated by different weight-updating frequencies, the following procedure is employed:

$$\begin{aligned}
&Bias_{\mathrm{ChLo}_2}(2011,8) - Bias_{\mathrm{ChLo}_3}(2011,8) \\
&= \left[P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{Target}}(2011,8)\right] - \left[P_{\mathrm{ChLo}_3}(2011,8) - P_{\mathrm{Target}}(2011,8)\right] \\
&= P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{ChLo}_3}(2011,8) \\
&= \left[P_{\mathrm{ChLo}_2}(2010,1) P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008})\right] - \left[P_{\mathrm{ChLo}_3}(2011,1) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009})\right] \\
&= P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000}) P_{\mathrm{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006}) \\
&\left\{ \begin{bmatrix} P_{\mathrm{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) P_{\mathrm{Lo}}(p^{2005,1}, p^{2006,1}, q^{2002}) P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004}) \\ P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,1}, q^{2008}) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2008}) \end{bmatrix} - \right. \\
&\left. \begin{bmatrix} P_{\mathrm{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000}) P_{\mathrm{Lo}}(p^{2005,1}, p^{2006,1}, q^{2003}) P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2003}) \\ P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,1}, q^{2006}) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009}) \end{bmatrix} \right\}
\end{aligned} \tag{21}$$

To facilitate the comparison, all the direct Lowe indexes in Equation (21) are written in terms of the indexes with the same price comparison periods. From the right-hand side of

Equation (21), it can be seen that the two pairs of Lowe indexes, $P_{\text{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000})$ and $P_{\text{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006})$, are identical; whereas, the other five pairs of Lowe indexes measure the price movement over the same periods but use different quantity weight vectors:

- In three pairs of Lowe indexes representing four years of price change – from January 2004 to January 2005 $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$, from January 2006 to January 2008 $P_{\text{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004})$ and $P_{\text{Lo}}(p^{2006,1}, p^{2008,1}, q^{2003})$, and from January 2010 to January 2011 $P_{\text{Lo}}(p^{2010,1}, p^{2011,1}, q^{2008})$ and $P_{\text{Lo}}(p^{2010,1}, p^{2011,1}, q^{2006})$ – those with a more frequent weight-updating schedule ($x = 2$) use relatively more up-to-date quantity weight vectors.
- Whereas, of the other two pairs of indexes corresponding to less than two years' price movement – one from January 2005 to January 2006, $P_{\text{Lo}}(p^{2005,1}, p^{2006,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2005,1}, p^{2006,1}, q^{2003})$, and the other from January 2011 to August 2011, $P_{\text{Lo}}(p^{2011,1}, p^{2011,8}, q^{2008})$ and $P_{\text{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009})$ – those with a less frequent weight-updating process ($x = 3$) use more up-to-date quantity weight vectors.

This simple comparison indicates that the chain-linked series with more frequent weight updates applies up-to-date quantity weights more often than those series with less frequent basket updates. Generally speaking, the price index compiled using a more outdated basket tends to exceed that which uses more up-to-date baskets due to price-induced commodity substitution. Thus, through this rough comparison, it is intuitively believed that more frequent weight updates would generate lower commodity-substitution bias in general.

To identify conditions under which more frequent weight updates would generate lower commodity-substitution bias, we compare one of the pairs of the Lowe indexes in Equation (21):

$$\begin{aligned} &P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) - P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000}) \\ &= \frac{\sum_i p_i^{2005,1} q_i^{2002}}{\sum_i p_i^{2004,1} q_i^{2002}} - \frac{\sum_i p_i^{2005,1} q_i^{2000}}{\sum_i p_i^{2004,1} q_i^{2000}} \\ &= \frac{\sum_i \left(\frac{p_i^{2005,1}}{p_i^{2004,1}} - P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) \right) \left(\frac{q_i^{2002}}{q_i^{2000}} - Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002}) \right) s_i^{2004,1:2000}}{Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})} \end{aligned} \tag{22}$$

where the Lowe quantity index, $Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})$, is defined as:

$$Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002}) = \frac{\sum_i p_i^{2004,1} q_i^{2002}}{\sum_i p_i^{2004,1} q_i^{2000}} \tag{23}$$

and the hybrid expenditure shares $s_i^{2004,1:2000}$ are defined in terms of the year 2000 quantity vector evaluated at January 2004 prices:

$$s_i^{2004,1:2000} = \frac{p_i^{2004,1} q_i^{2000}}{\sum_i p_i^{2004,1} q_i^{2000}} \tag{24}$$

The last line of Equation (22) indicates that the price deviations and quantity deviations are for two *different* periods; the former is pertaining to the period from January 2004 to January 2005, while the latter is for the period from year 2000 to 2002. Provided that the price and quantity changes were for the same period (e.g., from 2000 to 2002), the right-hand side of Equation (22) would be regarded as the covariance between the price deviations of price relatives from their mean, $\frac{p_i^{2002}}{p_i^{2000}} - P_{\text{Lo}}(p^{2000}, p^{2002}, q^{2002})$, and the corresponding quantity deviations of quantity relatives from their mean, $\frac{q_i^{2002}}{q_i^{2000}} - Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})$. If this covariance is negative (which is the usual case in the consumer context) and the price trend from 2000 to 2002 on average is in the same direction as those going from January 2004 to January 2005, the difference between the two Lowe indexes, shown in Equation (22), would be negative, which implies that the Lowe index using the up-to-date basket, $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$, will be lower than that using the out-dated basket, $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$.

In short, the relationship between $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$ depends upon the persistent tendency of price change and the associated change in consumers' expenditure patterns. This conclusion will also be true for the comparison of the other pairs of the Lowe indexes in Equation (21). However, the determination of the sign of Equation (21), which represents the relationship between the commodity-substitution biases in the Lowe indexes calculated with different frequencies of weight-updates, is far more complicated than what we have discussed here as it is affected by the interaction of the different time periods involved in the calculation. Despite this, from this simple example, we can still find that the impact of the frequency of weight updates on the upper-level commodity-substitution bias depends on the relationship between the price trend and the expenditure pattern of different time periods.

Intuitively, the more frequent the weights are updated, the more up-to-date weights would be employed in the index calculation. This is true for the comparison among other weight-updating frequencies. In addition, if persistent long-term price trends and consumers' price-induced commodity-substitution behaviour are present, then increasing the frequency of weight updates would lower the commodity-substitution bias.

4.1.2. Empirical Results: Impact of the Basket-Update Frequency on the Canadian CPI

Using the constructed data set, we compiled different CPI series by assuming different frequencies of updating the CPI basket while fixing the implementation lag equal to 13 months. Figure 2 shows the CPI series constructed with different frequencies of basket updates – from every year to every five years, and also with no basket updates at all, for the period from January 2002 to December 2013.

Series "Freq_x" ($x = 1, 2, \ldots 5$) in Figure 2 denotes CPI series compiled with the basket updated every x years. It illustrates that the index level for a given time period gradually decreases as the frequency of basket updates is accelerated. The index levels of

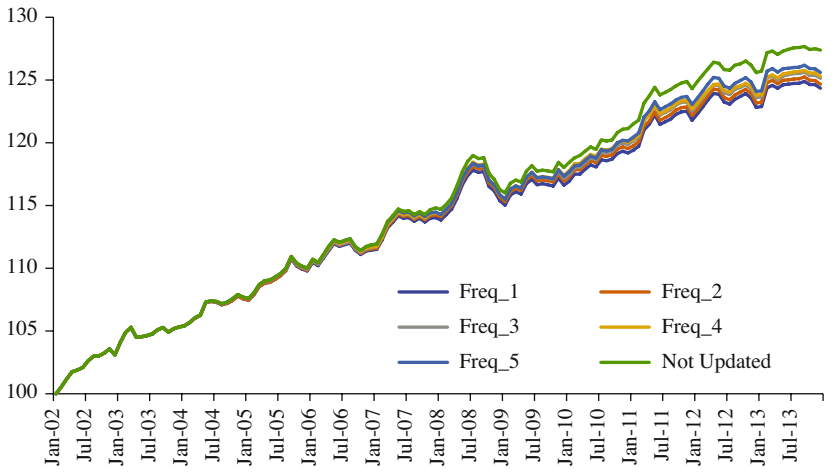

Fig. 2. Comparisons among the CPI series compiled with different frequencies of updating the CPI basket (January 2002 = 100).

the CPI series with no basket updates are considerably higher than levels of the other series. It is also noted that the differences in the index values are not obvious within the first five or six years. The impact of weight-updating frequency can be shown more explicitly in Table 2 in which the corresponding annual index levels were compared with the chain-linked Fisher index. The official CPI, compiled using a different data set, is not comparable to the other series reported in the table and is cited purely for reference.

Examining these results, we find that the commodity-substitution bias could be reduced by increasing the frequency of updating the CPI basket in the examined period; however, the magnitude of the marginal reduction in commodity-substitution bias for each additional increase in the frequency of basket updates varied. If we increased the frequency of updating the CPI basket from every two years to every year, we could reduce the commodity-substitution bias, measured by the difference of index growth rate, from

Table 2. Comparisons of different CPIs, compiled with various frequencies of basket updates and the Fisher index (2003–2011).

	Indexes (2003 = 100)	Difference in the indexes	Annual growth rate	Difference in the growth rate
	2003 to 2011		(%)	(%)
Fisher – *Target index*	114.389	0.000	1.695	0.000
Lowe index-every 1 year	115.857	1.468	1.857	0.162
Lowe index-every 2 years	116.153	1.764	1.889	0.195
Lowe index-every 3 years	116.547	2.157	1.932	0.238
Lowe index-every 4 years	116.645	2.256	1.943	0.249
Lowe index-every 5 years	116.918	2.528	1.973	0.278
Lowe index-no updates	118.009	3.620	2.091	0.397
Official CPI	116.70	2.3	1.944	0.249

0.195 percentage points to 0.162 percentage points on average. The impact was more significant when we changed the frequency from every four years to every two years, in which case the commodity-substitution bias was reduced from 0.249 percentage points to 0.195 percentage points on average for the sample period.

A similar impact on the CPI of increasing the frequency of weight updates was also shown in other studies, such as in Greenlees and Williams (2009) and in Ho et al. (2011). More recent research conducted by Australia, (Australian Bureau of Statistics 2016), found that the bias declined from 0.24% per year with six-year weight updates to 0.09% per year with one- year updates for the period between September 2005 and September 2011. Despite the magnitude of change being different between the two countries, we observe a similar impact of a reduced bias on the CPI through increasing the frequency of basket updates.

4.2. *Commodity-Substitution Bias and the Implementation Lag of a New Basket*

It is impossible to implement a new CPI basket in the weight reference period it refers to because of the time needed to conduct and process the Survey of Household Spending (SHS). This fact results in a certain time lag between the weight reference period and the implementation time of the basket. In this article, this time lag is referred to as the implementation lag. It is widely recognized that shortening the implementation lag of a new CPI basket can lower the upward bias in a Lowe price index. In this section, we will revisit this common belief and verify how this lag influences the CPI.

4.2.1. Conceptual Impact of the Implementation Lag on the CPI

If, for example, two baskets – the 2005 and 2009 baskets – are available for the period from January 2009 to December 2012, to implement the latter, we need a link month that chains indexes across the two baskets. To identify the impact of the implementation lag on the CPI, we assume that there are two possible link months, say December 2010 and April 2011, for introducing the 2009 basket. One has a shorter implementation lag (twelve months) while the other has a longer one (16 months). To assess the common belief in this simple setting, where a chain-linked Lowe index, defined in Equation (6), will be calculated, we compare the difference in the CPI series calculated using the two possible link months. Because of the inherent limitations of the Lowe formula, we believe that it will generate upward bias in most cases. Therefore, only upward bias will be taken into consideration.

For instance, the CPI from January 2009 to December 2012 using a shorter implementation lag, with December 2010 as the link month, denoted by $P_{\text{ChLo}}^{2010,12}(2012,12)$, can be compiled as follows:

$$\begin{aligned} P_{\text{ChLo}}^{2010,12}(2012,12) &= P_{\text{Lo}}(p^{2009,01}, p^{2010,12}, q^{2005})\, P_{\text{Lo}}(p^{2010,12}, p^{2012,12}, q^{2009}) \\ &= \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} \end{aligned} \tag{25}$$

The CPI for the same comparison periods using a longer lag, with April 2011 as the link month, denoted by $P_{\text{ChLo}}^{2011,04}(2012,12)$, can be compiled as follows:

$$\begin{aligned} P_{\text{ChLo}}^{2011,04}(2012,12) &= P_{\text{Lo}}(p^{2009,01}, p^{2011,04}, q^{2005})\, P_{\text{Lo}}(p^{2011,04}, p^{2012,12}, q^{2009}) \\ &= \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2011,04} q_i^{2009}} \end{aligned} \tag{26}$$

The difference in the magnitude of the commodity-substitution bias in the two CPIs can be derived from the following expression:

$$\begin{aligned} &\left[P_{\text{ChLo}}^{2010,12}(2012,12) - P_{\text{target}}(2012,12)\right] - \left[P_{\text{Ch-Lo}}^{2011,04}(2012,12) - P_{\text{target}}(2012,12)\right] \\ &= P_{\text{ChLo}}^{2010,12}(2012,12) - P_{\text{ChLo}}^{2011,04}(2012,12) \\ &= \left[\frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}}\right] - \left[\frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2011,04} q_i^{2009}} \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}}\right] \\ &= \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_i p_i^{2011,04} q_i^{2009}} \left(\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} - \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2010,12} q_n^{2005}}\right) \end{aligned} \tag{27}$$

A negative sign resulting from Equation (27) would imply that a shorter implementation lag leads to a lower commodity-substitution bias. Furthermore, the last line of Equation (27) indicates that the sign is determined by the difference between $\left(\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}}\right)$ and $\left(\frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2010,12} q_n^{2005}}\right)$, the two price indexes that measure price changes between the two link months (December 2010 and April 2011) with different baskets (the 2005 basket and 2009 basket). As mentioned before, generally speaking, price indexes using a more obsolete basket tend to exceed those using a more up-to-date basket due to consumers' substitution behaviour. If this is the case, the above difference would be negative, which leads to the conclusion that a shorter time lag would generate a lower bias as is commonly believed. However, is this intuition always true? To verify this, the difference between these two indexes is further examined.

To simplify the problem, we fix the products and services belonging to the two baskets. Decomposing the index difference yields the following expression:

$$\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} - \frac{\sum_i p_i^{2011,04} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}} = \sum_i \frac{\overbrace{\left(\frac{p_i^{2011,04}}{p_i^{2010,12}} - P_{\mathrm{Lo}}(p^{2010,12}, p^{2011,04}, q^{2009})\right)}^{\text{price deviation}} \overbrace{\left(\frac{q_i^{2009}}{q_i^{2005}} - Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009})\right)}^{\text{quantity deviation}}}{Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009})} s_i^{2010,12:2005} \tag{28}$$

where the Lowe quantity index is defined as:

$$Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009}) = \frac{\sum_i p_i^{2010,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2005}} \tag{29}$$

and the hybrid expenditure shares are defined as:

$$s_i^{2010,12:2005} = \frac{p_i^{2010,12} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}} \tag{30}$$

Thus, Equation (28) demonstrates that which link month yields lower commodity-substitution bias is determined by both price and quantity variations. It is, however, not easy to determine its sign, because the price and quantity deviations are for two different periods. If the deviations in both prices and quantities are for the same period, it could be regarded as the covariance between price relatives and the corresponding quantity relatives. In typical consumer theory, this covariance is negative – the price deviation $\left(\frac{p_i^{2009}}{p_i^{2005}} - P_{\mathrm{Lo}}(p^{2005}, p^{2009}, q^{2009})\right)$ and the quantity deviation $\left(\frac{q_i^{2009}}{q_i^{2005}} - Q_{\mathrm{Lo}}(p_i^{2010,12}, q^{2005}, q^{2009})\right)$ are negatively correlated. If the price trend between the two possible link months (December 2010 and April 2011), represented by $\left(\frac{p_i^{2011,04}}{p_i^{2010,12}} - P_{\mathrm{Lo}}(p^{2010,12}, p^{2011,04}, q^{2009})\right)$ is, on average, in the same direction as those between the two weight reference years (2005 and 2009), then we would expect that $\frac{\sum_i p_i^{2011,4} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}}$ exceeds $\frac{\sum_i p_i^{2011,4} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}}$. As a result, shortening the implementation lag could reduce the commodity-substitution bias.

In summary, this simplified case shows that a shorter implementation lag is associated with lower commodity-substitution bias as long as (i) the price trend between the two weight reference years is in the same direction as those between the two possible link months, and (ii) price-induced consumers' commodity-substitution behavior exists.

Price trends between the two weight reference years, in general, represent long-term price movements, whereas the price trends between two possible link months, if not too far from each other, normally reflect unpredictable price changes that are not necessarily in line with the long-term price movements, especially considering seasonal items. This implies that the impact on the CPI of shortening the implementation lag is not predictable.

It depends on the consistency between the long-term price trends and short-term price fluctuations, and on the presence of consumer's commodity-substitution behaviour. If prices of the majority of goods and services move persistently in the same direction for a long period, such as in an inflation context, this condition is more likely to be satisfied.

4.2.2. Empirical Results: Impact of the Implementation Lag on the Canadian CPI

In the first part of this section, we use the CPI series and apply the official CPI baskets without any adjustments, to examine whether shortening the implementation lag for introducing the 2005 basket, the 2009 basket, and the 2011 basket could reduce the commodity-substitution bias in the Canadian CPI.

The 2005 CPI basket was officially implemented in May 2007. Here we assume that it could have been implemented in any month from January 2007 to April 2007. Under operational constraints, we assume it is infeasible to implement the 2005 baskets earlier than January 2007. A negative difference would be shown in the fifth column of Table 3 if introducing the 2005 basket earlier than May 2007 could reduce the commodity-substitution bias. However, the numerical results reported in Table 3 imply that implementing the 2005 basket earlier than May 2007 would not yield a lower CPI bias.

Similarly the sign of the difference between $\frac{\sum_n p_n^{2011,4} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ and $\frac{\sum_n p_n^{2011,4} q_n^{2005}}{\sum_n p_n^{link} q_n^{2005}}$ listed in the fifth column of Table 4 would determine whether the commodity-substitution bias would have decreased or increased by introducing the 2009 basket earlier than May 2011. The sign of the difference between $\frac{\sum_n p_n^{2013,1} q_n^{2011}}{\sum_n p_n^{link} q_n^{2011}}$ and $\frac{\sum_n p_n^{2013,1} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ in the fifth column of Table 5 determines whether the commodity-substitution bias in the Canadian CPI could have decreased or increased by shortening the implementation lag of the 2011 basket. The results in both Table 4 and Table 5 show that reducing the implementation lag of the 2009 and 2011 baskets would not yield a lower CPI bias under the time constraints of the availability of the SHS.

In the following part of this section, the different CPI series calculated with different implementation lags using the constructed data set are reported. To isolate the impact of this phenomenon as opposed to the impact of weight-updating frequency, we fix the frequency of updating weights at every two years, and vary only the implementation lag to somewhere between twelve and 24 months. We also show the results of using one month as the implementation lag; although this is currently operationally impossible as the

Table 3. Different link months for introducing the 2005 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2007,4} q_n^{2005}}{\sum_n p_n^{link} q_n^{2005}}$ (A)	$\frac{\sum_i p_i^{2007,4} q_i^{2001}}{\sum_i p_i^{link} q_i^{2001}}$ (B)	Difference (A)−(B)
Jan. 2007	Dec. 2006	102.0116	101.9890	0.0226
Feb. 2007	Jan. 2007	101.9928	101.9447	0.0481
Mar. 2007	Feb. 2007	101.2472	101.2451	0.0021
Apr. 2007	Mar. 2007	100.3856	100.3813	0.0043

Table 4. Different link months for introducing the 2009 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2011,4} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ (A)	$\frac{\sum_i p_i^{2011,4} q_i^{2005}}{\sum_i p_i^{link} q_i^{2005}}$ (B)	Difference (A)−(B)
Jan. 2011	Dec. 2010	102.0339	102.0011	0.0329
Feb. 2011	Jan. 2011	101.7826	101.7540	0.0287
Mar. 2011	Feb. 2011	101.4966	101.4701	0.0266
Apr. 2011	Mar. 2011	100.4137	100.4076	0.0062

finalized expenditure data, taken mainly from the SHS, can be obtained only as early as eleven months after the weight reference year.

Figure 3 shows the cumulative impact of the implementation lag, which are kept unchanged for each CPI series, on the index values. In general, there are minor differences in index values when the implementation lags are not significantly different from each other; this explains why the ten CPI series cannot be distinguished separately in Figure 3. However, over time, the series with longer implementation lags clearly begin to diverge from the series with shorter lags (for example, 24 months compared to twelve months). It can also be demonstrated by the fact that the CPI series with a one-month implementation lag is significantly lower than the other CPI series.

Table 6 shows the comparison between the annual chained Fisher index and the annual chained Lowe price indexes compiled using different implementation lags for the period from 2003 to 2011, as well as the corresponding geometric average growth rates. The annual chained Lowe indexes are derived by taking a simple arithmetic average of monthly chained Lowe indexes of a calendar year. Even though the ways of calculating an annual chain Fisher index and annual chain Lowe index are different, the comparison still provides insights when comparing with the same target index. It clearly indicates how the index value and the average inflation rate change with the implementation lags. Among the chained-CPI series that can be compiled in a timely manner, using twelve months as the implementation lag yielded the lowest inflation rate; however, the difference in the average inflation rate between using twelve months and 14 months as the implementation lag was only 0.01 percentage points for the sample period. As expected from the conceptual framework, the impact of the implementation lag on the CPI is not predictable, especially when we shorten or increase the lags by increments of one or two months.

Table 5. Different link months for introducing the 2011 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2013,1} q_n^{2011}}{\sum_n p_n^{link} q_n^{2011}}$ (A)	$\frac{\sum_i p_i^{2013,1} q_i^{2009}}{\sum_i p_i^{link} q_i^{2009}}$ (B)	Difference (A)−(B)
Jan. 2013	Dec. 2012	100.0678	100.0567	0.0111
Feb. 2013	Jan. 2013	100.0000	100.0000	0.0000
Mar. 2013	Feb. 2013	98.8240	98.8067	0.0173

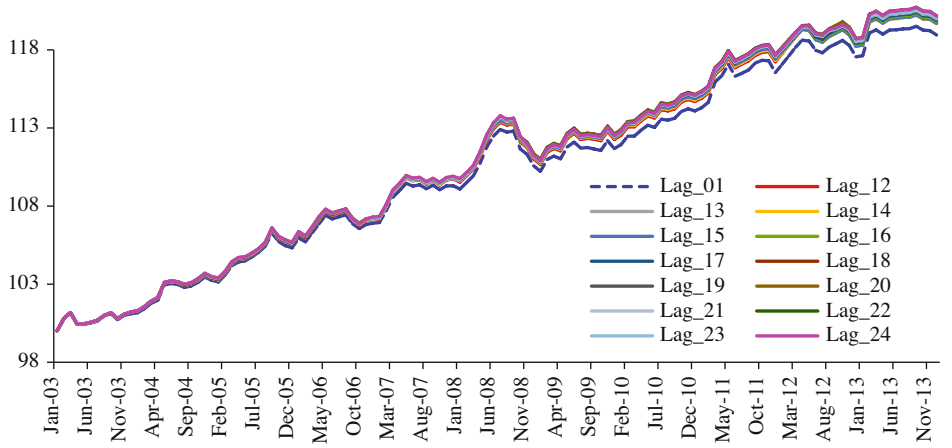

Fig. 3. Different CPI series corresponding to various implementations lags.

However, the commodity-substitution bias could generally be reduced if the implementation lag were substantially shortened. This can be shown from the difference in the growth rates between implementation lags of one month and twelve months, as well as twelve months and 24 months. Table 6 indicates that the substitution bias can be reduced from 0.221 percentage points to 0.176 percentage points if the implementation lag is shortened from 24 months to twelve months. ABS (2016) found a similar impact of the weight implementation lag on the CPI. The bias declined from 0.15% per year with a two-year implementation lag to 0.09% per year with a one-year implementation lag for the period from September 2005 to September 2011.

Table 6. Comparison of the geometric average growth rates of the different CPI series using various implementation lags and the Fisher index.

	Indexes (2003 = 100)	Differences in indexes	Annual growth rate	Difference in growth rate
	2003–2011		(%)	(%)
Fisher	114.389	0.000	1.695	0.000
Lowe index, 1 month lag	115.484	1.095	1.816	0.121
Lowe index, 12 month lag	115.980	1.591	1.870	0.176
Lowe index, 13 month lag	116.153	1.764	1.889	0.195
Lowe index, 14 month lag	116.075	1.686	1.881	0.186
Lowe index, 15 month lag	116.164	1.775	1.891	0.196
Lowe index, 16 month lag	116.300	1.911	1.905	0.211
Lowe index, 17 month lag	116.282	1.893	1.903	0.209
Lowe index, 18 month lag	116.340	1.951	1.910	0.215
Lowe index, 19 month lag	116.432	2.043	1.920	0.225
Lowe index, 20 month lag	116.413	2.023	1.918	0.223
Lowe index, 21 month lag	116.348	1.959	1.911	0.216
Lowe index, 22 month lag	116.405	2.016	1.917	0.222
Lowe index, 23 month lag	116.316	1.926	1.907	0.213
Lowe index, 24 month lag	116.393	2.004	1.916	0.221

Table 7. Different link months for introducing the 2010 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_m p_i^{2012,4} q_i^{2010}}{\sum_m p_i^{link} q_i^{2010}}$ (A) 2010 basket	$\frac{\sum_i p_i^{2012,4} q_i^{2008}}{\sum_i p_i^{link} q_i^{2008}}$ (B) 2008 basket	Difference (A)−(B)
Jan. 2012	Dec. 2011	101.707	101.589	0.118
Feb. 2012	Jan. 2012	101.263	101.192	0.071
Mar. 2012	Feb. 2012	100.821	100.773	0.048
Apr. 2012	Mar. 2012	100.382	100.370	0.011
May 2012	Apr. 2012	100.000	100.000	0.000
June 2012	May 2012	100.027	99.974	0.053
July 2012	June 2012	100.500	100.399	0.102
Aug. 2012	July 2012	100.629	100.470	0.159
Sep. 2012	Aug. 2012	100.338	100.167	0.171
Oct. 2012	Sep. 2012	100.170	100.083	0.087

From these empirical results, we cannot infer the impact on the CPI of a given link month of a particular CPI basket. To identify and illustrate this impact, we examine the introduction of a specific CPI basket. If, for example, the 2010 basket could be possibly implemented between January 2012 and October 2012, any month from December 2011 to September 2012 could, therefore, be chosen as the link month. Using Equation (28), we can determine retrospectively which month is the optimal link month for introducing the 2010 CPI basket. Table 7 shows the comparison between April 2012 and all the other possible link months, which are within the timeline of the SHS.

We obtained positive differences in the fifth column of Table 7, implying that using months either earlier or later than April 2012 as the link month cannot reduce commodity-substitution bias in the CPI based on Equation (28). Although using April 2012 as the link month to introduce the 2010 basket generates the lowest index level, it might not necessarily be true for introducing other new baskets. We therefore perform the same exercise (results are available on request) for the introduction of other baskets, and find that the optimal month for different baskets varies with the price fluctuation.

The empirical results illustrate that the impact of shortening the implementation lag on the commodity-substitution bias is not predictable, especially when the price trends are not persistent over time. However, in the case that a country's economy exhibited persistent and predictable inflation, the conditions implied by Equation (28) might very likely be satisfied. This could result in the observance of a relatively significant impact of a shortened implementation lag on the substitution bias.

Recently, as a result of operational constraints, Statistics Canada used 13 months as the implementation lag to introduce the 2011 basket. The empirical results from this study suggest that shortening the implementation lag to twelve months may not have a significant impact on further reducing the commodity-substitution bias. Moreover, the link month that yields the lowest commodity-substitution bias may not always be the same because of different monthly price fluctuations over time. As a result, it is not meaningful

to fix the link month of implementing a new basket for the purpose of reducing the commodity-substitution bias; in addition, the optimal link month of introducing a new CPI basket cannot be determined in advance. However, since Statistics Canada also compiles the CPI annual table based on the calendar year, we recommend that the new baskets be introduced in January to have a consistent annual index.

4.3. *Alternative Data Sources and Substitution Bias*

Many retailers, including nearly all major retailers, collect data through automated point-of-sale scanners. Scanner data is becoming an increasingly important source of information for statistical agencies, providing them with the prices and quantities of a large number of actual transactions in a timely manner. Several national statistical agencies currently make use of this data, including the Netherlands, Norway, Sweden, Switzerland, and New Zealand. Meanwhile, with the development of electronic commerce, online shopping has become more popular. Accompanying this growth, public information on product prices and characteristics is also available online. Automated data collection ("web-scraping") can replace traditional price collection for some product categories. With these "big data" sources, statistical agencies and academic researchers have an opportunity to study many research issues that used to be operationally infeasible and purely theoretical, and explore new methods to solve these issues.

With the availability of scanner data, it seems that the commodity-substitution bias issue raised in this article can more easily be addressed by using the prices and quantities to construct weighted (preferably superlative) price indexes. Research using scanner data to either estimate the substitution bias or to produce a superior estimate of the CPI has been ongoing for more than thirty years. New challenges and problems also arise with the use of scanner data, such as more volatile estimates of the CPI and chain-drift caused by the use of high-frequency scanner data. To overcome these new problems with the use of scanner data, Ivancic et al. (2011) proposed an innovative method, described as a rolling year GEKS method (RYGEKS), which adapts multilateral index number theory to making comparisons between multiple time periods. The GEKS method, described by Gini (1931), Eltetö and Köves (1964), and Szulc (1964), was originally used to conduct multilateral comparison, involving two stages of aggregation. The RYGEKS method makes maximum use of all matches in the scanner data to compile non-revisable CPIs that are approximately free from chain drift. Since then, this novel approach has been tested by many countries' numerical experiments, such as de Haan and van der Grient (2011) using Dutch data, Johansen and Nygaard (2011) using Norwegian data, and Krsinich (2011) using scanner data from New Zealand. Extensions to the RYGEKS have also been made; for instance, de Haan and Krisinich (2014) used an imputation Törnqvist rolling year GEKS procedure (ITRYGEKS) to derive quality-adjusted and chain-drift free price indexes. This method was applied by Statistics New Zealand (2014) to produce a CPI for its electronics products beginning in the September 2014 quarter.

Following the advancement of these new data processing techniques, the availability of large-volume data sources could provide statistical agencies with new practical solutions to primary questions covered in this paper. It might be feasible to obtain timely information on quantities purchased by households and to dramatically shorten the basket

implementation lag with the arrival of new data processing systems that could eliminate or supress some of the current operational constraints. However, more research is needed before incorporating these data sources into the CPI.

5. Conclusion

The Lowe index number formula, one of the fixed-basket concept indexes, is widely used by statistical agencies to compile their Consumer Price Index (CPI). However, because of its limitations associated with the fixed-basket concept, some concern arises from the use of this formula, in particular the issue of commodity-substitution bias. Because of the importance of the CPI to its different users (such as central banks, policy makers, and the general population as a whole), researchers have devoted, and continue to devote, much work into investigating the issue of commodity-substitution bias in the CPI.

In this article, we constructed a comprehensive data set by using information taken from Statistics Canada's Survey of Household Spending (SHS) for the years from 2000 to 2011, and the monthly CPI data for Canada at the basic class level for the period from January 2000 to December 2013.

This study focused on the investigation of approaches to reducing the commodity-substitution bias in the Canadian CPI based on two key aspects associated with the introduction of new CPI baskets. Namely, updating the CPI basket more frequently, and introducing a new CPI basket in a more timely manner. The empirical results found in this paper for the examination period indicate that increasing the frequency of updating the CPI basket could reduce the commodity-substitution bias. This finding is consistent with what has been shown in Greenlees and Williams (2009) and in other studies. In addition this paper's results reveal that the marginal gains from moving from basket-updates every four years to every two years are more significant than those from moving from basket-updates every two years to every year.

The impact of shortening the implementation lag for a new CPI basket on the commodity-substitution bias is unpredictable because it depends on the consistency between the long-term price trends (between the two basket reference periods) and the short-term price movements (between the possible link months), as well as the existence of consumers' price-induced commodity-substitution behaviour. Clear differences can be perceived in the price indexes compiled by using a twelve-month implementation lag versus an 18-month or longer implementation lag, while the differences in the indexes are largely reduced when they are constructed using twelve months compared to 14 months as implementation lags. Therefore, based on both the decomposition of index differences and the empirical results in this article, it is believed that the conclusion in Généreux (1983) would hold only when the conditions illustrated above were satisfied. Consequently, it is worthwhile for a statistical agency to pursue ways to dramatically shorten the implementation lag; however, taking great effort to slightly improve the timeliness of implementing a new basket may not provide meaningful returns.

In this article, we presented the empirical results using Canadian data for the period between 2003 and 2011. These results do not provide direct answers for choosing the most effective approach to reducing the commodity-substitution bias in a CPI. Statistical agencies in other countries can draw inferences from this empirical work but should be

cautious in generalizing these results to other CPIs because of the time dependence of the empirical results. Finally, new practical solutions associated with the incorporation of large-volume data sources in the CPI is worth further investigation from statistical agencies.

6. References

Australian Bureau of Statistics. 2016. "Increasing the Frequency of Consumer Price Index Expenditures Class Weight Updates." *Australian Bureau of Statistics Information Paper* 6401.0.60.002. July.

Bérubé, C. 1996. "Selecting a Formula for the Canadian CPI, 1962–1994." Price Division Analytical Series, Statistics Canada. No. 7, Catalogue No. 62F0014MPB, p 7.

Boskin, Michael J., Ellen R. Dullberger, Robert J. Gordon, Zvi Griliches, and Dale W. Jorgenson. 1996. "Toward a More Accurate Measure of the Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index." Washington, D.C.: US Government Printing Office.

Crawford, A. 1998. "Measurement Biases in the Canadian CPI: An Update." *Bank of Canada Review*. Spring. Page 39–56.

De Haan, Jan and Heymerik A. van der Grient. 2011. "Eliminating Chain Drift in Price Indexes Based on Scanner Data." *Journal of Econometrics* 161(1): 36–46.

De Haan, Jan and Frances Krisnich. 2014. "Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes." *Journal of Business & Economic Statistics* 32(3): 341–358.

Diewert, W. Erwin. 1976. "Exact and Superlative Index Numbers." *Journal of Econometrics.* May. Page 115–145.

Eltetö, Ö. and P. Köves. 1964. "On a Problem of Index Number Computation Relating to International Comparisons." *Statisztikai Szemle* 42: 507–518. (in Hungarian).

Généreux, Pierre A. 1983. "Impact of the Choice of Formulae on the Canadian Consumer Price Index." *Price Level Measurement: Proceedings from a Conference* Sponsored by Statistics Canada. Erwin Diewert and Claude Montmarquette (eds.).

Gini, C. 1931. "On the Circular Test of Index Numbers." *Metron* 9(9): 3–24.

Greenlees, John S. and Elliot Williams. 2009. *Reconsideration of Weighting and Updating Procedure in the US CPI*. Paper presented at the 11th Meeting of the International Working Group on Price Indexes (Ottawa group) in Neuchâtel, Switzerland.

Ho, Ricky, Peter Campion, and Chris Pike. 2011. *New Zealand Consumer Price Index—an empirical analysis of the frequency and level of weight updates*. Room document at the 12th Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand.

ILO / IMF / OECD / IMECE / Eurostat / World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO.

Ivancic, Lorraine, W. Erwin Diewert, and Kevin J. Fox. 2011. "Scanner Data, Time Aggregation and Construction of Price Indexes." *Journal of Econometrics* 161(1): 24–35.

Johansen, Ingvild, and Ragnhild Nygaard. 2011. *Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages*. Paper presented at the 12th

Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand.

Krsinich, Frances. 2011. *Price indexes from scanner data*. Paper presented at the 12th Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand. May 2011.

Rossiter, James. 2005. *Measurement bias in the Canadian Consumer Price Index*. Bank of Canada Working Paper 2005-39. December.

Sabourin, Patrick. 2012. "Measurement bias in the Canadian Consumer Price Index: An update." *Bank of Canada Review*. Summer.

Shiratsuka, Shigenori. 2006. "Measurement Errors in Japanese Consumer Price Index." *Monetary and Economic Studies* 17(3). Institute for Monetary and Economic Studies, Bank of Japan, 69–102.

Statistics Canada. 2014. "The Canadian Consumer Price Index Reference Paper." Catalogue No. 62-553-X.

Statistics New Zealand. 2014. *Measuring price change for consumer electronics using scanner data*. Available from www.stats.govt.nz.

Szulc, B. 1964. "Indexes for Multiregional Comparisons." *Przeglad Statystyczny* 3: 239–254. (in Polish).

Wynne, Mark A. and Diego Rodriguez-Palenzuela. 2002. *Measurement Bias in the HICP: What do we know, and what do we need to know?* European Central Bank Working Paper Series, No. 131. Frankfurt: European Central Bank.

Received January 2016
Revised October 2016
Accepted January 2017

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 979–1004, http://dx.doi.org/10.1515/JOS-2017-0046

The Effects of the Frequency and Implementation Lag of Basket Updates on the Canadian CPI

Ning Huang[1], *Waruna Wimalaratne*[1], *and Brent Pollard*[1]

In this article, we examine the effects of different frequencies and implementation months of basket updates on the fixed-basket price index – the Lowe index, through theoretical analysis and empirical simulation using Canadian data from 2000 to 2013. We find that both an increased frequency of basket updates and a faster implementation of these new baskets will reduce substitution bias in the CPI. However, we also find that improvements to the method of accelerating frequency has diminishing marginal returns in practice – as each subsequent increase in the frequency with which the CPI basket is updated has a less pronounced effect; and the ideal link-month when a new basket is implemented is unpredictable, since the impact of the implementation lag depends upon the consistency between short-term price movements and long-term price trends.

Key words: Consumer price index; Lowe formula; fixed-basket index; basket-update frequency; implementation lag; measurement bias; commodity substitution bias; superlative indexes; Fisher, Walsh, Törnqvist indexes.

1. Introduction

The Consumer Price Index (CPI) is the most widely used indicator of price change in Canada. It serves a variety of purposes, and is therefore of interest to governments, unions, business organizations, research institutions, and the general public. Its various uses include its function as a general indicator of inflation in Canada and as a tool for adjusting incomes, wages, and other payments to ensure that purchasing power is unaffected by any average price movement. Further details are available in a reference paper published by Statistics Canada (2014).

In line with the practices of most other national statistical agencies, the Consumer Prices Division (CPD) at Statistics Canada uses the Lowe index formula for aggregating its CPI at the upper level. The Lowe index formula, often described as a "Laspeyres-type" formula, is a fixed-basket formula. This means that the quantity and quality of the goods

[1] Statistics Canada, Consumer Prices Division, 170 Tunney's Pasture, Ottawa, O.N., K1A 0T6. Canada. Emails: Nings.Huang@Canada.ca, Waruna.Wimalaratne@Canada.ca, and Brent.Pollard@Canada.ca.

Acknowledgments: The authors thank Ross Beck-MacNeil, Nathalie Brault, Kyle de March, Xin Ha, Mathieu Lequain, Sue Morris, Marc Prud'homme, Bradley Snider, Philip Smith, Faouzi Tarkhani, Amanda Wright, Alice Xu, Clément Yélou, two internal reviewers at Statistics Canada, and Price Measurement Advisory Committee members Jan de Haan, Erwin Diewert, Pierre Duguay, David Fenwick, John Mallon, Marshal Reinsdorf, Patrick Sabourn, Mick Silver, Kam Yu for their helpful discussions and comments. Thanks also to three anonymous reviewers for their comments. The views expressed in this article are solely those of the authors and do not necessarily reflect those of Statistics Canada.

and services included in the CPI basket must be unchanged or equivalent within the life span of a CPI basket. It is also referred to as a Cost of Goods Index (COGI).

The Lowe formula is used in practice because it offers a simple and convenient way to compile composite price indexes in a timely manner. Although the Lowe formula is a good choice for the fixed-basket concept of a CPI, its inherent limitations must be taken into consideration; for example, it cannot account for consumer's price-induced product substitution, it experiences delay in reflecting the effects of new goods and services on consumer price change, and it has difficulty in fully accounting for changes in the quality of existing consumer products. Due to these and other limitations, the official CPI, published by Statistics Canada, is not a true measure of actual changes in the cost of living.

A Cost-of-Living Index (COLI) is derived from the standpoint of an economic theory, based upon the assumption of a household's utility optimization behaviour, which assumes that a household will structure its purchases to maximize utility, or satisfaction, given a certain level of prices and a certain level of income. Since a household's utility optimization problem is dual to its cost minimization problem, a COLI then measures the change in the household's minimum cost of maintaining a fixed level of utility over two periods when faced with changes in prices. The theory of the COLI provides the conceptual framework for some countries' CPI, such as the United States (U.S.) and Sweden.

The difference between the official CPI and an underlying COLI, which can be approximated by a class of superlative indexes, is called measurement bias. According to the Consumer Price Index (CPI) Manual (ILO et al. 2004), a group of "superlative" price indexes, such as the Fisher, Walsh and Törnqvist indexes, is expected to provide "fairly close" approximations to the underlying COLI. Thus, they are recommended in the manual as the "target indexes" for the upper-level index. The main types of measurement bias include commodity-substitution bias, outlet-substitution bias, quality-change bias and new-goods bias. In this study, only the measurement bias associated with the upper-level aggregation is discussed. Apart from this bias, there could also be sampling and other non-sampling bias in the estimated elementary indexes and estimated basket weights. Note that measurement bias can be measured in terms of index level and index growth rate. In this article, commodity-substitution bias is analyzed and reported in both ways depending on the context of the article.

These measurement biases arise from the fact that any basket weights, held constant over more than one period, do not necessarily reflect the types of purchases that consumers actually make to attain the same level of welfare when relative prices change. A fixed-basket index, therefore, normally fails to account for the changes in consumers' purchasing patterns or preferences in a timely manner, and measures only the average price movement based on a specifically defined basket, resulting in measurement bias.

A COLI, on the other hand, allows for changes in the basket over time and, therefore, accounts for changes in consumer purchasing patterns when measuring average price movements over two periods. While numerous national statistical offices do not construct their CPIs as a COLI, including Statistics Canada, many of them still want to have knowledge about the measurement bias in their official CPI because of its important role as a major economic indicator and as a wage or salary indexation factor.

Since the CPI is the most commonly used indicator for tracking overall price change in Canada, measurement bias in the CPI is an important issue for both its users and compilers.

As Sabourin (2012) pointed out, "since the CPI departs from a true COLI, it is subject to measurement bias and does not necessarily reflect changes in the wellbeing of consumers, which could be problematic for monetary policy and when making cost-of-living adjustments to wages and salaries."

Given the varying uses of the CPI, research on the measurement bias in the Canadian CPI is conducted regularly by some of its users, such as continuous research conducted by the Bank of Canada, including Crawford (1998), Rossiter (2005) and Sabourin (2012). According to Sabourin (2012), for the years from 2005 to 2011, the mean total bias in the Canadian CPI was 0.45 percentage points per year from 2005 to 2011, among which commodity-substitution bias was 0.22, outlet substitution bias was 0.04, new-good bias was 0.20 and quality adjustment bias was −0.01. Similar studies quantifying the bias in the CPI have been conducted in other countries, such as the paper by Boskin et al. (1996), also known as the Boskin Commission Report, for the U.S., which stated that "the Commission's best estimate of the size of the upward bias looking forward is 1.1 percentage points per year. The range of plausible value is 0.8 to 1.6 percentage points per year." The estimates of CPI bias can also be found in Shiratsuka (2006) for Japan and in Wynne and Rodriguez-Palenzuela (2002) for European countries.

This article focuses on the investigation of commodity-substitution bias, which is caused by the inability of a fixed-basket index to capture consumers' price-induced substitution. Generally speaking, without changing the formula for compiling the CPI, this type of bias could be reduced by updating the CPI basket more frequently and by implementing the basket in a more timely fashion. Both of these methods allow a more accurate reflection of the changes in purchasing patterns due to consumers' substitution between different combinations of goods and services. In the existing literature associated with commodity-substitution bias, there are only a limited number of studies examining the impact on the CPI of the frequency and delay of implementing new basket weights. This is likely due to the difficulties associated with acquiring such data. In the Canadian context, the annual household expenditure survey facilitated this study.

It is widely recognized that more frequent basket updates and faster implementation will lead to an index that more closely approximates a superlative measure. For instance, Japan publishes two series of CPI: the official CPI, with weights updated every five years; and a chained Laspeyres CPI, with weights updated annually.

A study by Greenlees and Williams (2009) showed that quarterly weight updates generated an index that more closely resembled a target index when compared to less frequent updates. In their study, a chained Törnqvist index was calculated as a superlative target. They simulated various weight updating periods: quarterly, semi-annual, annual and biennial. The index derived from quarterly weights approximated most closely to the superlative index. They also found that the Lowe index updated annually, which could be realistically compiled under the operational constraints, increased less than the rolling, two-year index of current methodology in four out of the six years studied (2002 to 2007). In addition, the advantage of using more timely weights was not offset by any increase in index volatility or instability.

Ho et al. (2011) examined, using data from 2002 to 2008, the impact on the New Zealand CPI of reweighting at different frequencies and at different levels of the index structure. They showed that frequent weight updates at the sub-item level and above generated CPI

series that tracked the Fisher series most closely among those generated by using other weight-update frequencies and other aggregation levels. Their current methodology with weight updates in June 2002, 2006, and 2008 quarters yielded a Laspeyres index of 117.0, while their methodology without updates produced an index of 117.9; these can be compared to a Fisher of 115.8, for the June 2008 quarter.

In addition to the frequency of basket updates, national statistical offices also need to determine when to introduce a new basket. The delay in the implementation of a new basket affects the size of commodity-substitution bias. Limited research supports this: Généreux (1983), using Canadian data, compared a chained Laspeyres series with eight basket updates against a chained Laspeyres series with only one basket update over the period from 1957 to 1978. He concluded "what appears to be desirable is not necessarily a more frequent updating of the CPI baskets but a more timely one." For example, implementing the new weights in the years they refer to could considerably reduce the commodity-substitution bias. Using Canadian data, Bérubé (1996) also showed that introducing a basket two years after the basket reference period would reduce the annual substitution bias from 0.20 percentage points to 0.18 percentage points over the period from 1962 to 1994, compared with introducing a basket three years after the reference period.

A study from Australia Bureau Statistics (ABS 2016) showed a significant decrease in substitution bias by having shorter weight implementation lag for the period between September 2005 and September 2011. The bias declined from 0.24% per year for the CPI to 0.09%, 0.15%, and 0.16% with weight implementation lags of one, two, and three years, respectively. The Australian CPI weights are updated every six years using a household survey. In their study they utilized household final consumption expenditure from National Accounts to calculate the Lowe Index.

In 2010, Statistics Canada implemented the *CPI Enhancement Initiative*, a multi-stage program to advance the quality of the CPI. As part of this initiative, effort was directed at identifying and reducing the commodity-substitution bias. In 2013, a more frequent basket update schedule was implemented – from once every four years to once every two years. Additionally, the 2011 basket was introduced more quickly than past baskets – the time lag went from 16 months to 13 months. Interest and focus subsequently shifted to investigating the effect that changes such as these have on the quality of the CPI. The results would help inform the decision of whether to further accelerate the frequency of basket updates and further reduce the implementation lag.

The Canadian economy, similar to those of other major economies, is a knowledge-based economy, associated with dynamic technological change. With the rapid applications of new technology and emergence of new products and new market structure, consumers' lifestyles and merchants' pricing strategies have also experienced significant change. As a result, it is expected that a CPI basket becomes outdated more rapidly.

This in turn, raises questions for compilers of CPIs attempting to improve index quality and accuracy: does the comparison by Généreux (1983) between "a more frequent updating of the CPI baskets" and "a more timely one" still hold? Are empirical results from other countries, such as those revealed by Greenlees and Williams (2009) also valid for Canada? And, how can national statistical offices reduce the commodity substitution bias further?

Updating the basket weights of a price index such as the CPI is accomplished in various stages. How each of these are implemented will likely have some effect on the overall

index. This article will focus on the performance of the index under different scenarios for two of these stages, the weight-updating cycle, and the timeliness of the introduction of the new weights. The principal source of the data for the study is the Canadian Survey of Household Spending, which is used to reflect changes in consumers' spending patterns over time, for the period from 2002 to 2013. Price indexes from the Canadian CPI are also used.

To estimate the substitution bias, this article compares the results of the Lowe price index with those of the Fisher price index. This approach differs from the more common method of estimation, which compares the results between the Laspeyres price index and the Fisher price index. Another difference lies in the focus of the analysis: instead of only reporting the empirical results derived from Canadian data, the divergence in the resulting indexes obtained under various scenarios (different weight-updating schedules, and different implementation lags of the introduction of new weights) is analyzed in detail using a mathematical approach. Consequently, this article will shed new light on how to mitigate the well-known and pervasive substitution bias which characterizes a fixed-basket CPI for national statistical offices in countries facing similar situations as Canada. For these countries, the findings will play an important role in determining the desired frequency of weight updates and time of implementing new weights.

The remainder of this article is organized as follows: Section 2 discusses the data sources and data construction methods; Section 3 defines the target price index, which belongs to a group of superlative series that closely approximate a COLI, used in this study; Section 4 addresses the effects of the frequency and implementation lag of weight updates on the Canadian CPI in detail; and Section 5 concludes the article.

2. Data Construction

The two main elements required for the calculation of a price index series are prices and quantities. To this end, this study makes use of two main sources of data – the Consumer Price Index (CPI) and the Survey of Household Spending (SHS). The CPI provides data on the price indexes for each of its measured goods and services at the basic class level of aggregation. Basic classes are the lowest-level aggregates of products, chosen by Statistics Canada, for which a set of weights is fixed for the duration of the CPI basket. The SHS data are used in constructing fixed-basket weights for twelve years going from 2000 to 2011 based on the 2005 CPI classification structure. In this way, the estimated substitution bias would not be affected by the impact of changes in the specification and the appearance of new products.

The "price" component of the index calculation comes from the CPI over the period from January 2000 to December 2013. The original price indexes are unlinked price indexes for each of the corresponding published CPI basket. To facilitate the index reconstruction, the indexes were linked together based on the classification of the 2005 basket and rebased to January 2000 = 100. The reconstructed indexes, therefore, represent the price movement from the price reference period of January 2000 to a given price observation month.

The "quantity" component of the index comes from the SHS, which contains detailed information about consumer spending during a given reference year. The SHS sample has

a cross-sectional design, and is selected from the Labour Force Survey sampling frame and carried out in private households. The SHS is the main source of the expenditure weights data for the CPI.

In the first stage of the data construction, we derived expenditure weights for the years without official CPI weights – 2000, 2002 to 2004, 2006 to 2008, and 2010, using data from the SHS. Official CPI weights data were used whenever they were available; specifically the 2001, 2005, 2009, and 2011 baskets. However, some adjustments were made in order to align them with the 2005 classification of the CPI at the basic class level of aggregation. The 2005 classification structure that was in use in the official CPI from May 2007 to April 2011 was maintained across time to preserve uniformity and avoid complications arising from the introduction of new items. For non-official basket update years, some expenditure values were unavailable from the SHS; for example, the low level details for some basic classes under the food classification. To estimate these expenditure values, we used a modified price-updating method which used a weighted average of expenditures for those years with detailed SHS information. With this method, relatively greater importance is assigned to expenditures in baskets from periods closer to the imputed period. For example, the unknown expenditure for item i in 2003 can be imputed from the formula:

$$p_i^{2003}q_i^{2003} \equiv (6/8)\underbrace{p_i^{2001}q_i^{2001}\frac{p_i^{2003}}{p_i^{2001}}}_{\text{upward price update}} + (2/8)\underbrace{p_i^{2009}q_i^{2009}\frac{p_i^{2003}}{p_i^{2009}}}_{\text{backward price update}} \quad (1)$$

Finally, similar imputation strategies were employed for calculating the weights for the mortgage interest cost basic class as well as some components of the clothing classification. In the case of the mortgage interest cost index, where Statistics Canada has a special treatment, data were available only for the official basket reference years. As a remedy, weights for the remaining years were calculated using the same method as employed for food classification. For the replacement cost basic class, the SHS lacked detailed housing data for non-official basket update years, and so a combination of internal and external data was used to calculate its value.

Once the "price" and "quantity" components were built, a data validation was performed by reconstructing the official CPI using the analytical database. Comparing the constructed CPI with the official CPI, we believe that the analytical series was a very good approximation.

3. Target Index Formula

To determine the magnitude of the commodity-substitution bias, first, it is necessary to select a target index with which to compare the estimates of this study. The Fisher, Walsh, and Törnqvist indexes have been widely used for this purpose, as they belong to a small class of "superlative indexes".

An important characteristic of superlative indexes is that they include the prices and quantities in both periods being compared, they are therefore symmetrically weighted indexes. Moreover, these three index number formulas are flexible and provide second-order approximation to each other. In other words, different superlative indexes tend to

have similar properties, yield similar results and behave in very similar ways. In addition, they are expected to provide a close approximation to the underlying conditional cost-of-living index (COLI). Diewert (1976) showed that superlative indexes provide close approximations to any true cost-of-living price index if the underlying utility function is linear homogeneous. As a close approximation to the unknown COLI, superlative indexes are recommended in the CPI ILO Manual as the theoretical target indexes. The difference between the Laspeyres-type index, which does not permit the commodity-substitution induced by relative price changes, and the target indexes can be treated as a measure of commodity-substitution bias at the upper level of index aggregation when holding classification structure unchanged.

In this study, we aim at comparing chained-CPI series constructed by applying different weights. The target indexes are, therefore, estimated by using the chain-linked Fisher, Walsh, and Törnqvist index number formulas with annual weight-updating, as detailed monthly expenditure data are unavailable. The corresponding annual CPI series are derived by taking the unweighted arithmetic average of monthly price indexes of the twelve months in the calendar year. Using the Fisher index number formula $P_{\mathrm{ChF}}^{(2003+t)/2003}$ as an example, we show how the chain-linked index between 2003 and 2011 is constructed:

$$
\begin{aligned}
P_{\mathrm{ChF}}^{(2003+t)/2003} &= \prod_{j=1}^{t} P_{\mathrm{F}}^{(2003+j)/(2003+j-1)} \\
&= \prod_{j=1}^{t} \left(\underbrace{\frac{\sum_{i=1}^{N} p_i^{2003+j} q_i^{2003+j-1}}{\sum_{i=1}^{N} p_i^{2003+j-1} q_i^{2003+j-1}}}_{P_{\mathrm{L}}^{2003+j/2003+j-1}} \underbrace{\frac{\sum_{i=1}^{N} p_i^{2003+j} q_i^{2003+j}}{\sum_{i=1}^{N} p_i^{2003+j-1} q_i^{2003+j}}}_{P_{\mathrm{P}}^{2003+j/2003+j-1}} \right)^{1/2} \qquad (2) \\
& \quad t = 1, 2, \ldots 8
\end{aligned}
$$

where $P_{\mathrm{ChF}}^{(2003+t)/2003}$ denotes chained-Fisher from 2003 to 2003 + t; $P_{\mathrm{F}}^{(2003+j)/(2003+j-1)}$ denotes the direct Fisher index from 2003 + j − 1 to 2003 + j; $P_{\mathrm{L}}^{(2003+j)/(2003+j-1)}$ denotes the direct Laspeyres index from 2003 + j − 1 to 2003 + j and $P_{\mathrm{P}}^{(2003+j)/(2003+j-1)}$ denotes the direct Paasche index from 2003 + j − 1 to 2003 + j. N is the total number of goods and services included in the CPI basket. The chained-Walsh index and chained-Törnqvist index can be compiled similarly. The three superlative indexes are expected to behave similarly, which is confirmed by the numerical results over the period from 2003 to 2011 reported in Table 1, where the average growth rate using the chained Fisher index as an example, is calculated as $\sqrt[8]{\left(P_{\mathrm{ChF}}^{2011/2003}/100\right)} - 1$.

In the next section, the target index values in Table 1 can be compared with the CPI series compiled with different CPI weight-updating schedules to produce the estimates of the upper-level commodity-substitution bias. More specifically, the chained-Fisher index is used as an example to estimate the commodity-substitution bias in this article.

Table 1. Superlative price indexes (2003 = 100).

Year (2003 + t)	Fisher		Walsh		Törnqvist	
	Chained index	Annual inflation	Chained index	Annual inflation	Chained index	Annual inflation
2003	100.000		100.000		100.000	
2004	101.728	1.728	101.730	1.730	101.730	1.730
2005	103.746	1.984	103.750	1.986	103.750	1.986
2006	105.475	1.667	105.480	1.668	105.482	1.669
2007	107.401	1.826	107.409	1.829	107.410	1.828
2008	109.624	2.069	109.632	2.070	109.633	2.069
2009	109.670	0.042	109.684	0.047	109.688	0.050
2010	111.404	1.581	111.422	1.585	111.422	1.581
2011	114.389	2.679	114.408	2.680	114.405	2.677
Average growth rate (2003–2011)		1.695		1.697		1.696

4. Approaches to Reducing Commodity-Substitution Bias

In general, the commodity-substitution bias could be measured as the difference between the published CPI and the target index, both of which are estimated by keeping the items in the baskets fixed over time. The source of this substitution bias varies. Two important sources could be the frequency of CPI basket weight updates, and the time lag between the end of the basket reference year and the initial implementation time of a new CPI basket in the CPI calculation. Using the 2011 basket update as an example, we illustrate the relationship among different time periods involved in the index calculation using the following timeline.

On the timeline in Figure 1, the basket reference year (during which the SHS is conducted to collect the necessary information for the CPI basket) is 2011. The 2011 CPI basket was implemented with the February 2013 CPI, which is defined as the implementation month in this paper. The duration from January 2012 to January 2013 is the implementation lag, which in this case is 13 months. January 2013 is the link month for the implementation of the 2011 CPI basket.

In this section, how the frequency and implementation lag of the CPI weight affect the magnitude of the upper-level commodity-substitution bias will be explored in further detail.

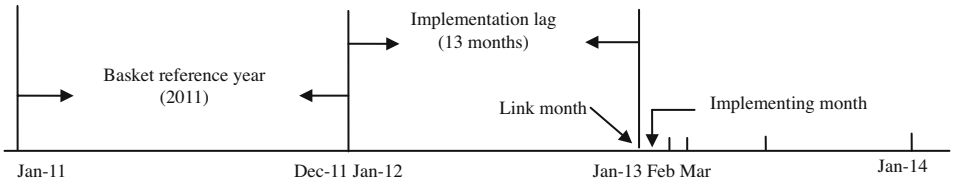

Fig. 1. Timeline of CPI basket update.

4.1. Commodity-Substitution Bias and the Frequency of Basket Updates

4.1.1. Conceptual Framework to Measure the Impact of the Basket Update Frequency

The CPI basket is designed to reflect consumers' spending patterns. As a result of both relative price changes and some long-term effects on consumers' spending behaviour, such as the impact of demographic factors and technological changes, the weights might become out-of-date and less representative of current consumption patterns. The bias in a Lowe index is likely to increase as the basket weights age. Therefore, CPI weights should be updated periodically to reflect the changes in these patterns.

To identify the pure impact of the frequency of weight updates on the magnitude of the CPI bias, we fix the implementation lag at 13 months and vary only the frequency of weight updates when calculating the All-items CPI, which measures price change of all the goods and services included in the Canadian CPI, for the period from January 2002 to December 2013.

A direct Lowe index $P_{\mathrm{Lo}}(p^0, p^t, q^b)$ can be defined in terms of a quantity vector $q^b \equiv \left[q_1^b, \ldots, q_N^b\right]$, a price vector of base period $p^0 \equiv \left[p_1^0, \ldots, p_N^0\right]$ and a price vector of current period $p^t \equiv \left[p_1^t, \ldots, p_N^t\right]$:

$$P_{\mathrm{Lo}}(p^0, p^t, q^b) = \frac{\sum_{i=1}^{N} p_i^t q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} \tag{3}$$

where N is the total number of goods and services included in the CPI weight structure.

It can be also written in terms of the hybrid share form as follows:

$$\begin{aligned} P_{\mathrm{Lo}}(p^0, p^t, s^{0:b}) &= \frac{\sum_{i=1}^{N} p_i^t q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} = \sum_{i=1}^{N} \left(\frac{p_i^t}{p_i^0}\right) \frac{p_i^0 q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b} \\ &= \sum_{i=1}^{N} \left(\frac{p_i^t}{p_i^0}\right) s_i^{0:b} \end{aligned} \tag{4}$$

where the hybrid expenditure shares $s_i^{0:b}$ corresponding to the quantity weights vector q^b measured at base period price vector p^0 are defined as:

$$s_i^{0:b} = \frac{p_i^0 q_i^b}{\sum_{i=1}^{N} p_i^0 q_i^b}, \quad i = 1, 2, \ldots, N \tag{5}$$

If more than one basket, say baskets $b1$ and $b2$, are in use, it is necessary to calculate the chain-linked Lowe index, where the indexes calculated using different CPI baskets are linked together. To explain this concept, let $p^{y,m}$ be the elementary price vector for year $y \geq 2002$ and month $m = 1, 2, \ldots, 12$; the chain-linked Lowe index for year y and month m, with every x years as the frequency of weight updates, is denoted as $P_{\mathrm{ChLo}_x}(y, m)$. The calculation of the chain-linked Lowe index depends on which basket is currently used and in which month it is linked to the previous basket. In general,

a chain-linked Lowe index can be defined as:

$$P_{\mathrm{ChLo}_x}(y,m) = P_{\mathrm{ChLo}_x}(link\_month)\, P_{\mathrm{Lo}}(p^{link\_month}, p^{y,m}, q^{b}) \tag{6}$$

where $P_{\mathrm{ChLo}_x}(link\_month)$ is a chain-linked Lowe index for the link month that chains together indexes using the current basket q^{b} and the previous baskets.

If the CPI basket is assumed to be updated every x years, where x can be 1, 2, 3, 4, or 5, after the adoption of the 2000 basket, the Equation (6) can be applied to compile the CPI series. With the implementation lag set equal to 13 months, a new basket $2000 + kx$ is introduced in February of year $2002 + kx$ with January ($m = 1$) of year $2002 + kx$ as the link month, $k = 1, 2, \ldots$ such that $2002 + kx \leq y$ (y is the year of the price index). With these assumptions, the chain-linked Lowe index can be calculated by substituting the corresponding values in Equation (6), yielding the following results:

$$P_{\mathrm{ChLo}_x}(y,m) = P_{\mathrm{ChLo}_x}(2002 + kx, 1)\, P_{\mathrm{Lo}}(p^{2002+kx,1}, p^{y,m}, q^{2000+kx}) \tag{7}$$

The first component of the right-hand side of the Equation (7), $P_{\mathrm{ChLo}_x}(2002 + kx, 1)$, is the link factor, which is also a chain-linked Lowe index for January of year $2002 + kx$, which is the link month of the current basket ($2000 + kx$); the second component, $P_{\mathrm{Lo}}(p^{2002+kx,1}, p^{y,m}, q^{2000+kx})$, is the direct Lowe index comparing the current month (y, m) with the link month ($2002 + kx$, *1*), January of year $2002 + kx$.

The link factor $P_{\mathrm{ChLo}_x}(2002 + kx, 1)$ can be also defined as the product of several direct Lowe indexes as follows:

$$\begin{aligned} P_{\mathrm{ChLo}_x}(2002 + kx, 1) = {} & P_{\mathrm{Lo}}(p^{0}, p^{2002+x,1}, q^{2000}) \\ & P_{\mathrm{Lo}}(p^{2002+x,1}, p^{2002+2x,1}, q^{2000+x}) \\ & \cdots P_{\mathrm{Lo}}(p^{2002+(k-1)x,1}, p^{2002+kx,1}, q^{2000+(k-1)x}) \end{aligned} \tag{8}$$

where k denotes the number of times the CPI basket is updated since the price reference period, which is assumed to be during the life span of the basket q^{2000}.

We now describe how the chain-linked Lowe index can be constructed if the weights are updated every two years, that is $x = 2$. Denote the chain-linked Lowe index for year y and month m, with an update frequency of every two years, by $P_{\mathrm{ChLo}_2}(y,m)$. In this case, the direct Lowe index, which uses the 2000 basket only, is employed from February 2002 to January 2004, with January 2002 as the link month, the overlapping period that links the old and new CPI series. Applying (9), we have $P_{\mathrm{ChLo}_2}(2002, 1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2002,1}, q^{2000}) = 1$. Thus, the chain-linked Lowe index defined by (9) is, for the first 24 months running from February 2002 to January 2004, equal to the direct Lowe index:

$$\begin{aligned} P_{\mathrm{ChLo}_2}(y,m) = {} & P_{\mathrm{Lo}}(p^{2002,1}, p^{y,m}, q^{2000}) \\ & (\text{with } y = 2002,\ 2003;\ \text{and } m = 1, 2, \ldots, 12 \text{ and } y = 2004;\ m = 1) \end{aligned} \tag{9}$$

The same direct Lowe index on the right hand side of (9) is, therefore, used to define the chain-linked Lowe index for January 2004:

$$P_{\mathrm{ChLo}_2}(2004,1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000}) \tag{10}$$

The above chain-linked Lowe index for January 2004 corresponds to the link factor that chains together indexes using the 2000 basket and the 2002 basket. For the remaining months in 2004 and 2005, the annual quantity weights vector q^{2002} becomes available and the chain-linked Lowe index is defined as follows:

$$\begin{gathered} P_{\mathrm{ChLo}_2}(y,m) = P_{\mathrm{ChLo}_2}(2004,1)\, P_{\mathrm{Lo}}(p^{2004,1}, p^{y,m}, q^{2002}) \\ (\text{with } y = 2004,\ 2005;\ m = 1,2,\ldots,12;\ y = 2006;\ m = 1) \end{gathered} \tag{11}$$

The chain-linked Lowe index for January 2006 is, therefore, defined as follows:

$$P_{\mathrm{ChLo}_2}(2006,1) = P_{\mathrm{ChLo}_2}(2004,1)\, P_{\mathrm{Lo}}(p^{2004,1}, p^{2006,1}, q^{2002}) \tag{12}$$

Here again, the chain-linked Lowe index for January 2006 is the link factor that chains indexes based on 2004, 2002, and 2000 baskets respectively. From February 2006 to January 2008, the annual quantity weights vector q^{2004} becomes available and the chain-linked Lowe for this time span is defined as follows:

$$\begin{gathered} P_{\mathrm{ChLo}_2}(y,m) = P_{\mathrm{ChLo}_2}(2006,1)\, P_{\mathrm{Lo}}(p^{2006,1}, p^{y,m}, q^{2004}) \\ (\text{with } y = 2006,\ 2007;\ m = 1,2,\ldots,12;\ y = 2008;\ m = 1) \end{gathered} \tag{13}$$

Once more, the link factor chaining the indexes together across baskets is the chain-linked Lowe index for January 2008 which continues to be defined by the right-hand side of (13), as follows:

$$P_{\mathrm{ChLo}_2}(2008,1) = P_{\mathrm{ChLo}_2}(2006,1)\, P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004}) \tag{14}$$

Continuing the above process, we can construct the chain-linked Lowe index for other months in the other years.

To show how the defined process works, here we compile a chain-linked Lowe index for a particular month, say August 2011, as an example. Assume the weight-updating frequency is two ($x = 2$) and implementation lag is 13 months. The chained Lowe index is then denoted by $P_{\mathrm{ChLo}_2}(2011,8)$. Based on the described process, the current period, August 2011, is identified to be in the time span going from February 2010 to January 2012 and the associated quantity weights vector is q^{2008}, with January 2010 as the link month. The chain-linked Lowe index $P_{\mathrm{ChLo}_2}(2011,8)$ can then be constructed as:

$$P_{\mathrm{ChLo}_2}(2011,8) = P_{\mathrm{ChLo}_2}(2010,1)\, P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008}) \tag{15}$$

where $P_{\mathrm{ChLo}_2}(2010,1)$ is the link factor that chains together the price indexes using the 2008 basket and the previous baskets. Based on Equation (8), it can be written as a product

of direct Lowe indexes as follows:

$$P_{\mathrm{ChLo}_2}(2010,1) = P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000})\, P_{\mathrm{Lo}}(p^{2004,1}, p^{2006,1}, q^{2002}) \\ P_{\mathrm{Lo}}(p^{2006,01}, p^{2008,1}, q^{2004})\, P_{\mathrm{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006}) \tag{16}$$

The direct Lowe index on the right-hand side of (15) can be compiled based on Equation (3) as follows:

$$P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008}) = \frac{\sum_i p_i^{2011,8}\, q_i^{2008}}{\sum_i p_i^{2010,1}\, q_i^{2008}} \tag{17}$$

Next, the chain-linked Lowe index for the same month, August 2011, but with different weight-updating frequency, $x = 3$, denoted by $P_{\mathrm{ChLo}_3}(2011,8)$, is considered. It can be compiled based on the process described in the case of a weight update every two years (refer to Equation (9) to (14)), as follows:

$$P_{\mathrm{ChLo}_3}(2011,8) = P_{\mathrm{ChLo}_3}(2011,1)\, P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009}) \tag{18}$$

With the two CPI index values associated with different frequencies of weight updates, the commodity-substitution bias can be then estimated by comparing the chain-linked Lowe index with the same target index. For example, let $Bias_{\mathrm{ChLo}_2}(2011,8)$ and $Bias_{\mathrm{ChLo}_3}(2011,8)$ denote the commodity-substitution bias, measured in terms of index level, of the chain-linked Lowe index for August 2011, with weight-updating frequencies equal to every two and every three years, respectively. They can be defined as follows

$$Bias_{\mathrm{ChLo}_2}(2011,8) = P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{Target}}(2011,8) \tag{19}$$

$$Bias_{\mathrm{ChLo}_3}(2011,8) = P_{\mathrm{ChLo}_3}(2011,8) - P_{\mathrm{Target}}(2011,8) \tag{20}$$

To compare the magnitude of the bias generated by different weight-updating frequencies, the following procedure is employed:

$$\begin{aligned}
&Bias_{\mathrm{ChLo}_2}(2011,8) - Bias_{\mathrm{ChLo}_3}(2011,8) \\
&= \left[P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{Target}}(2011,8)\right] - \left[P_{\mathrm{ChLo}_3}(2011,8) - P_{\mathrm{Target}}(2011,8)\right] \\
&= P_{\mathrm{ChLo}_2}(2011,8) - P_{\mathrm{ChLo}_3}(2011,8) \\
&= \left[P_{\mathrm{ChLo}_2}(2010,1) P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,8}, q^{2008})\right] - \left[P_{\mathrm{ChLo}_3}(2011,1) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009})\right] \\
&= P_{\mathrm{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000}) P_{\mathrm{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006}) \\
&\left\{ \begin{bmatrix} P_{\mathrm{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) P_{\mathrm{Lo}}(p^{2005,1}, p^{2006,1}, q^{2002}) P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004}) \\ P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,1}, q^{2008}) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2008}) \end{bmatrix} - \right. \\
&\left. \begin{bmatrix} P_{\mathrm{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000}) P_{\mathrm{Lo}}(p^{2005,1}, p^{2006,1}, q^{2003}) P_{\mathrm{Lo}}(p^{2006,1}, p^{2008,1}, q^{2003}) \\ P_{\mathrm{Lo}}(p^{2010,1}, p^{2011,1}, q^{2006}) P_{\mathrm{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009}) \end{bmatrix} \right\}
\end{aligned} \tag{21}$$

To facilitate the comparison, all the direct Lowe indexes in Equation (21) are written in terms of the indexes with the same price comparison periods. From the right-hand side of

Equation (21), it can be seen that the two pairs of Lowe indexes, $P_{\text{Lo}}(p^{2002,1}, p^{2004,1}, q^{2000})$ and $P_{\text{Lo}}(p^{2008,1}, p^{2010,1}, q^{2006})$, are identical; whereas, the other five pairs of Lowe indexes measure the price movement over the same periods but use different quantity weight vectors:

- In three pairs of Lowe indexes representing four years of price change – from January 2004 to January 2005 $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$, from January 2006 to January 2008 $P_{\text{Lo}}(p^{2006,1}, p^{2008,1}, q^{2004})$ and $P_{\text{Lo}}(p^{2006,1}, p^{2008,1}, q^{2003})$, and from January 2010 to January 2011 $P_{\text{Lo}}(p^{2010,1}, p^{2011,1}, q^{2008})$ and $P_{\text{Lo}}(p^{2010,1}, p^{2011,1}, q^{2006})$ – those with a more frequent weight-updating schedule ($x = 2$) use relatively more up-to-date quantity weight vectors.
- Whereas, of the other two pairs of indexes corresponding to less than two years' price movement – one from January 2005 to January 2006, $P_{\text{Lo}}(p^{2005,1}, p^{2006,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2005,1}, p^{2006,1}, q^{2003})$, and the other from January 2011 to August 2011, $P_{\text{Lo}}(p^{2011,1}, p^{2011,8}, q^{2008})$ and $P_{\text{Lo}}(p^{2011,1}, p^{2011,8}, q^{2009})$ – those with a less frequent weight-updating process ($x = 3$) use more up-to-date quantity weight vectors.

This simple comparison indicates that the chain-linked series with more frequent weight updates applies up-to-date quantity weights more often than those series with less frequent basket updates. Generally speaking, the price index compiled using a more outdated basket tends to exceed that which uses more up-to-date baskets due to price-induced commodity substitution. Thus, through this rough comparison, it is intuitively believed that more frequent weight updates would generate lower commodity-substitution bias in general.

To identify conditions under which more frequent weight updates would generate lower commodity-substitution bias, we compare one of the pairs of the Lowe indexes in Equation (21):

$$\begin{aligned} &P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) - P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000}) \\ &= \frac{\sum_i p_i^{2005,1} q_i^{2002}}{\sum_i p_i^{2004,1} q_i^{2002}} - \frac{\sum_i p_i^{2005,1} q_i^{2000}}{\sum_i p_i^{2004,1} q_i^{2000}} \\ &= \frac{\sum_i \left(\frac{p_i^{2005,1}}{p_i^{2004,1}} - P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002}) \right) \left(\frac{q_i^{2002}}{q_i^{2000}} - Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002}) \right) s_i^{2004,1:2000}}{Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})} \end{aligned} \tag{22}$$

where the Lowe quantity index, $Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})$, is defined as:

$$Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002}) = \frac{\sum_i p_i^{2004,1} q_i^{2002}}{\sum_i p_i^{2004,1} q_i^{2000}} \tag{23}$$

and the hybrid expenditure shares $s_i^{2004,1:2000}$ are defined in terms of the year 2000 quantity vector evaluated at January 2004 prices:

$$s_i^{2004,1:2000} = \frac{p_i^{2004,1} q_i^{2000}}{\sum_i p_i^{2004,1} q_i^{2000}} \tag{24}$$

The last line of Equation (22) indicates that the price deviations and quantity deviations are for two *different* periods; the former is pertaining to the period from January 2004 to January 2005, while the latter is for the period from year 2000 to 2002. Provided that the price and quantity changes were for the same period (e.g., from 2000 to 2002), the right-hand side of Equation (22) would be regarded as the covariance between the price deviations of price relatives from their mean, $\frac{p_i^{2002}}{p_i^{2000}} - P_{\text{Lo}}(p^{2000}, p^{2002}, q^{2002})$, and the corresponding quantity deviations of quantity relatives from their mean, $\frac{q_i^{2002}}{q_i^{2000}} - Q_{\text{Lo}}(p^{2004,1}, q^{2000}, q^{2002})$. If this covariance is negative (which is the usual case in the consumer context) and the price trend from 2000 to 2002 on average is in the same direction as those going from January 2004 to January 2005, the difference between the two Lowe indexes, shown in Equation (22), would be negative, which implies that the Lowe index using the up-to-date basket, $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$, will be lower than that using the out-dated basket, $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$.

In short, the relationship between $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2002})$ and $P_{\text{Lo}}(p^{2004,1}, p^{2005,1}, q^{2000})$ depends upon the persistent tendency of price change and the associated change in consumers' expenditure patterns. This conclusion will also be true for the comparison of the other pairs of the Lowe indexes in Equation (21). However, the determination of the sign of Equation (21), which represents the relationship between the commodity-substitution biases in the Lowe indexes calculated with different frequencies of weight-updates, is far more complicated than what we have discussed here as it is affected by the interaction of the different time periods involved in the calculation. Despite this, from this simple example, we can still find that the impact of the frequency of weight updates on the upper-level commodity-substitution bias depends on the relationship between the price trend and the expenditure pattern of different time periods.

Intuitively, the more frequent the weights are updated, the more up-to-date weights would be employed in the index calculation. This is true for the comparison among other weight-updating frequencies. In addition, if persistent long-term price trends and consumers' price-induced commodity-substitution behaviour are present, then increasing the frequency of weight updates would lower the commodity-substitution bias.

4.1.2. Empirical Results: Impact of the Basket-Update Frequency on the Canadian CPI

Using the constructed data set, we compiled different CPI series by assuming different frequencies of updating the CPI basket while fixing the implementation lag equal to 13 months. Figure 2 shows the CPI series constructed with different frequencies of basket updates – from every year to every five years, and also with no basket updates at all, for the period from January 2002 to December 2013.

Series "Freq_x" ($x = 1, 2, \ldots 5$) in Figure 2 denotes CPI series compiled with the basket updated every x years. It illustrates that the index level for a given time period gradually decreases as the frequency of basket updates is accelerated. The index levels of

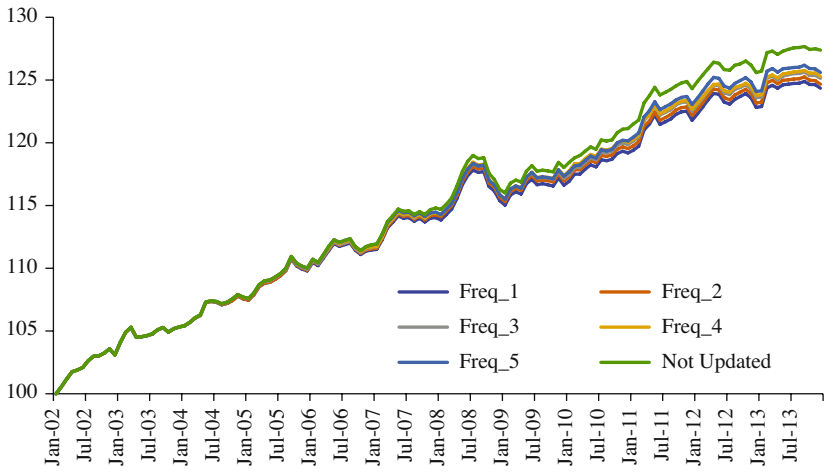

Fig. 2. Comparisons among the CPI series compiled with different frequencies of updating the CPI basket (January 2002 = 100).

the CPI series with no basket updates are considerably higher than levels of the other series. It is also noted that the differences in the index values are not obvious within the first five or six years. The impact of weight-updating frequency can be shown more explicitly in Table 2 in which the corresponding annual index levels were compared with the chain-linked Fisher index. The official CPI, compiled using a different data set, is not comparable to the other series reported in the table and is cited purely for reference.

Examining these results, we find that the commodity-substitution bias could be reduced by increasing the frequency of updating the CPI basket in the examined period; however, the magnitude of the marginal reduction in commodity-substitution bias for each additional increase in the frequency of basket updates varied. If we increased the frequency of updating the CPI basket from every two years to every year, we could reduce the commodity-substitution bias, measured by the difference of index growth rate, from

Table 2. Comparisons of different CPIs, compiled with various frequencies of basket updates and the Fisher index (2003–2011).

	Indexes (2003 = 100)	Difference in the indexes	Annual growth rate	Difference in the growth rate
	2003 to 2011		(%)	(%)
Fisher – *Target index*	114.389	0.000	1.695	0.000
Lowe index-every 1 year	115.857	1.468	1.857	0.162
Lowe index-every 2 years	116.153	1.764	1.889	0.195
Lowe index-every 3 years	116.547	2.157	1.932	0.238
Lowe index-every 4 years	116.645	2.256	1.943	0.249
Lowe index-every 5 years	116.918	2.528	1.973	0.278
Lowe index-no updates	118.009	3.620	2.091	0.397
Official CPI	116.70	2.3	1.944	0.249

0.195 percentage points to 0.162 percentage points on average. The impact was more significant when we changed the frequency from every four years to every two years, in which case the commodity-substitution bias was reduced from 0.249 percentage points to 0.195 percentage points on average for the sample period.

A similar impact on the CPI of increasing the frequency of weight updates was also shown in other studies, such as in Greenlees and Williams (2009) and in Ho et al. (2011). More recent research conducted by Australia, (Australian Bureau of Statistics 2016), found that the bias declined from 0.24% per year with six-year weight updates to 0.09% per year with one- year updates for the period between September 2005 and September 2011. Despite the magnitude of change being different between the two countries, we observe a similar impact of a reduced bias on the CPI through increasing the frequency of basket updates.

4.2. *Commodity-Substitution Bias and the Implementation Lag of a New Basket*

It is impossible to implement a new CPI basket in the weight reference period it refers to because of the time needed to conduct and process the Survey of Household Spending (SHS). This fact results in a certain time lag between the weight reference period and the implementation time of the basket. In this article, this time lag is referred to as the implementation lag. It is widely recognized that shortening the implementation lag of a new CPI basket can lower the upward bias in a Lowe price index. In this section, we will revisit this common belief and verify how this lag influences the CPI.

4.2.1. Conceptual Impact of the Implementation Lag on the CPI

If, for example, two baskets – the 2005 and 2009 baskets – are available for the period from January 2009 to December 2012, to implement the latter, we need a link month that chains indexes across the two baskets. To identify the impact of the implementation lag on the CPI, we assume that there are two possible link months, say December 2010 and April 2011, for introducing the 2009 basket. One has a shorter implementation lag (twelve months) while the other has a longer one (16 months). To assess the common belief in this simple setting, where a chain-linked Lowe index, defined in Equation (6), will be calculated, we compare the difference in the CPI series calculated using the two possible link months. Because of the inherent limitations of the Lowe formula, we believe that it will generate upward bias in most cases. Therefore, only upward bias will be taken into consideration.

For instance, the CPI from January 2009 to December 2012 using a shorter implementation lag, with December 2010 as the link month, denoted by $P_{\text{ChLo}}^{2010,12}(2012,12)$, can be compiled as follows:

$$\begin{aligned} P_{\text{ChLo}}^{2010,12}(2012,12) &= P_{\text{Lo}}(p^{2009,01}, p^{2010,12}, q^{2005})\, P_{\text{Lo}}(p^{2010,12}, p^{2012,12}, q^{2009}) \\ &= \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} \end{aligned} \tag{25}$$

The CPI for the same comparison periods using a longer lag, with April 2011 as the link month, denoted by $P_{\text{ChLo}}^{2011,04}(2012,12)$, can be compiled as follows:

$$\begin{aligned} P_{\text{ChLo}}^{2011,04}(2012,12) &= P_{\text{Lo}}(p^{2009,01}, p^{2011,04}, q^{2005})\, P_{\text{Lo}}(p^{2011,04}, p^{2012,12}, q^{2009}) \\ &= \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2011,04} q_i^{2009}} \end{aligned} \tag{26}$$

The difference in the magnitude of the commodity-substitution bias in the two CPIs can be derived from the following expression:

$$\begin{aligned} &\left[P_{\text{ChLo}}^{2010,12}(2012,12) - P_{\text{target}}(2012,12)\right] - \left[P_{\text{Ch-Lo}}^{2011,04}(2012,12) - P_{\text{target}}(2012,12)\right] \\ &= P_{\text{ChLo}}^{2010,12}(2012,12) - P_{\text{ChLo}}^{2011,04}(2012,12) \\ &= \left[\frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}}\right] - \left[\frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_i p_i^{2011,04} q_i^{2009}} \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2009,01} q_n^{2005}}\right] \\ &= \frac{\sum_i p_i^{2012,12} q_i^{2009}}{\sum_n p_n^{2009,01} q_n^{2005}} \frac{\sum_n p_n^{2010,12} q_n^{2005}}{\sum_i p_i^{2011,04} q_i^{2009}} \left(\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} - \frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2010,12} q_n^{2005}}\right) \end{aligned} \tag{27}$$

A negative sign resulting from Equation (27) would imply that a shorter implementation lag leads to a lower commodity-substitution bias. Furthermore, the last line of Equation (27) indicates that the sign is determined by the difference between $\left(\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}}\right)$ and $\left(\frac{\sum_n p_n^{2011,04} q_n^{2005}}{\sum_n p_n^{2010,12} q_n^{2005}}\right)$, the two price indexes that measure price changes between the two link months (December 2010 and April 2011) with different baskets (the 2005 basket and 2009 basket). As mentioned before, generally speaking, price indexes using a more obsolete basket tend to exceed those using a more up-to-date basket due to consumers' substitution behaviour. If this is the case, the above difference would be negative, which leads to the conclusion that a shorter time lag would generate a lower bias as is commonly believed. However, is this intuition always true? To verify this, the difference between these two indexes is further examined.

To simplify the problem, we fix the products and services belonging to the two baskets. Decomposing the index difference yields the following expression:

$$\frac{\sum_i p_i^{2011,04} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}} - \frac{\sum_i p_i^{2011,04} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}} = \sum_i \frac{\overbrace{\left(\frac{p_i^{2011,04}}{p_i^{2010,12}} - P_{\mathrm{Lo}}(p^{2010,12}, p^{2011,04}, q^{2009})\right)}^{\text{price deviation}} \overbrace{\left(\frac{q_i^{2009}}{q_i^{2005}} - Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009})\right)}^{\text{quantity deviation}}}{Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009})} s_i^{2010,12:2005} \tag{28}$$

where the Lowe quantity index is defined as:

$$Q_{\mathrm{Lo}}(p^{2010,12}, q^{2005}, q^{2009}) = \frac{\sum_i p_i^{2010,12} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2005}} \tag{29}$$

and the hybrid expenditure shares are defined as:

$$s_i^{2010,12:2005} = \frac{p_i^{2010,12} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}} \tag{30}$$

Thus, Equation (28) demonstrates that which link month yields lower commodity-substitution bias is determined by both price and quantity variations. It is, however, not easy to determine its sign, because the price and quantity deviations are for two different periods. If the deviations in both prices and quantities are for the same period, it could be regarded as the covariance between price relatives and the corresponding quantity relatives. In typical consumer theory, this covariance is negative – the price deviation $\left(\frac{p_i^{2009}}{p_i^{2005}} - P_{\mathrm{Lo}}(p^{2005}, p^{2009}, q^{2009})\right)$ and the quantity deviation $\left(\frac{q_i^{2009}}{q_i^{2005}} - Q_{\mathrm{Lo}}(p_i^{2010,12}, q^{2005}, q^{2009})\right)$ are negatively correlated. If the price trend between the two possible link months (December 2010 and April 2011), represented by $\left(\frac{p_i^{2011,04}}{p_i^{2010,12}} - P_{\mathrm{Lo}}(p^{2010,12}, p^{2011,04}, q^{2009})\right)$ is, on average, in the same direction as those between the two weight reference years (2005 and 2009), then we would expect that $\frac{\sum_i p_i^{2011,4} q_i^{2005}}{\sum_i p_i^{2010,12} q_i^{2005}}$ exceeds $\frac{\sum_i p_i^{2011,4} q_i^{2009}}{\sum_i p_i^{2010,12} q_i^{2009}}$. As a result, shortening the implementation lag could reduce the commodity-substitution bias.

In summary, this simplified case shows that a shorter implementation lag is associated with lower commodity-substitution bias as long as (i) the price trend between the two weight reference years is in the same direction as those between the two possible link months, and (ii) price-induced consumers' commodity-substitution behavior exists.

Price trends between the two weight reference years, in general, represent long-term price movements, whereas the price trends between two possible link months, if not too far from each other, normally reflect unpredictable price changes that are not necessarily in line with the long-term price movements, especially considering seasonal items. This implies that the impact on the CPI of shortening the implementation lag is not predictable.

It depends on the consistency between the long-term price trends and short-term price fluctuations, and on the presence of consumer's commodity-substitution behaviour. If prices of the majority of goods and services move persistently in the same direction for a long period, such as in an inflation context, this condition is more likely to be satisfied.

4.2.2. Empirical Results: Impact of the Implementation Lag on the Canadian CPI

In the first part of this section, we use the CPI series and apply the official CPI baskets without any adjustments, to examine whether shortening the implementation lag for introducing the 2005 basket, the 2009 basket, and the 2011 basket could reduce the commodity-substitution bias in the Canadian CPI.

The 2005 CPI basket was officially implemented in May 2007. Here we assume that it could have been implemented in any month from January 2007 to April 2007. Under operational constraints, we assume it is infeasible to implement the 2005 baskets earlier than January 2007. A negative difference would be shown in the fifth column of Table 3 if introducing the 2005 basket earlier than May 2007 could reduce the commodity-substitution bias. However, the numerical results reported in Table 3 imply that implementing the 2005 basket earlier than May 2007 would not yield a lower CPI bias.

Similarly the sign of the difference between $\frac{\sum_n p_n^{2011,4} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ and $\frac{\sum_n p_n^{2011,4} q_n^{2005}}{\sum_n p_n^{link} q_n^{2005}}$ listed in the fifth column of Table 4 would determine whether the commodity-substitution bias would have decreased or increased by introducing the 2009 basket earlier than May 2011. The sign of the difference between $\frac{\sum_n p_n^{2013,1} q_n^{2011}}{\sum_n p_n^{link} q_n^{2011}}$ and $\frac{\sum_n p_n^{2013,1} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ in the fifth column of Table 5 determines whether the commodity-substitution bias in the Canadian CPI could have decreased or increased by shortening the implementation lag of the 2011 basket. The results in both Table 4 and Table 5 show that reducing the implementation lag of the 2009 and 2011 baskets would not yield a lower CPI bias under the time constraints of the availability of the SHS.

In the following part of this section, the different CPI series calculated with different implementation lags using the constructed data set are reported. To isolate the impact of this phenomenon as opposed to the impact of weight-updating frequency, we fix the frequency of updating weights at every two years, and vary only the implementation lag to somewhere between twelve and 24 months. We also show the results of using one month as the implementation lag; although this is currently operationally impossible as the

Table 3. Different link months for introducing the 2005 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2007,4} q_n^{2005}}{\sum_n p_n^{link} q_n^{2005}}$ (A)	$\frac{\sum_i p_i^{2007,4} q_i^{2001}}{\sum_i p_i^{link} q_i^{2001}}$ (B)	Difference (A)−(B)
Jan. 2007	Dec. 2006	102.0116	101.9890	0.0226
Feb. 2007	Jan. 2007	101.9928	101.9447	0.0481
Mar. 2007	Feb. 2007	101.2472	101.2451	0.0021
Apr. 2007	Mar. 2007	100.3856	100.3813	0.0043

Table 4. Different link months for introducing the 2009 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2011,4} q_n^{2009}}{\sum_n p_n^{link} q_n^{2009}}$ (A)	$\frac{\sum_i p_i^{2011,4} q_i^{2005}}{\sum_i p_i^{link} q_i^{2005}}$ (B)	Difference (A)−(B)
Jan. 2011	Dec. 2010	102.0339	102.0011	0.0329
Feb. 2011	Jan. 2011	101.7826	101.7540	0.0287
Mar. 2011	Feb. 2011	101.4966	101.4701	0.0266
Apr. 2011	Mar. 2011	100.4137	100.4076	0.0062

finalized expenditure data, taken mainly from the SHS, can be obtained only as early as eleven months after the weight reference year.

Figure 3 shows the cumulative impact of the implementation lag, which are kept unchanged for each CPI series, on the index values. In general, there are minor differences in index values when the implementation lags are not significantly different from each other; this explains why the ten CPI series cannot be distinguished separately in Figure 3. However, over time, the series with longer implementation lags clearly begin to diverge from the series with shorter lags (for example, 24 months compared to twelve months). It can also be demonstrated by the fact that the CPI series with a one-month implementation lag is significantly lower than the other CPI series.

Table 6 shows the comparison between the annual chained Fisher index and the annual chained Lowe price indexes compiled using different implementation lags for the period from 2003 to 2011, as well as the corresponding geometric average growth rates. The annual chained Lowe indexes are derived by taking a simple arithmetic average of monthly chained Lowe indexes of a calendar year. Even though the ways of calculating an annual chain Fisher index and annual chain Lowe index are different, the comparison still provides insights when comparing with the same target index. It clearly indicates how the index value and the average inflation rate change with the implementation lags. Among the chained-CPI series that can be compiled in a timely manner, using twelve months as the implementation lag yielded the lowest inflation rate; however, the difference in the average inflation rate between using twelve months and 14 months as the implementation lag was only 0.01 percentage points for the sample period. As expected from the conceptual framework, the impact of the implementation lag on the CPI is not predictable, especially when we shorten or increase the lags by increments of one or two months.

Table 5. Different link months for introducing the 2011 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_n p_n^{2013,1} q_n^{2011}}{\sum_n p_n^{link} q_n^{2011}}$ (A)	$\frac{\sum_i p_i^{2013,1} q_i^{2009}}{\sum_i p_i^{link} q_i^{2009}}$ (B)	Difference (A)−(B)
Jan. 2013	Dec. 2012	100.0678	100.0567	0.0111
Feb. 2013	Jan. 2013	100.0000	100.0000	0.0000
Mar. 2013	Feb. 2013	98.8240	98.8067	0.0173

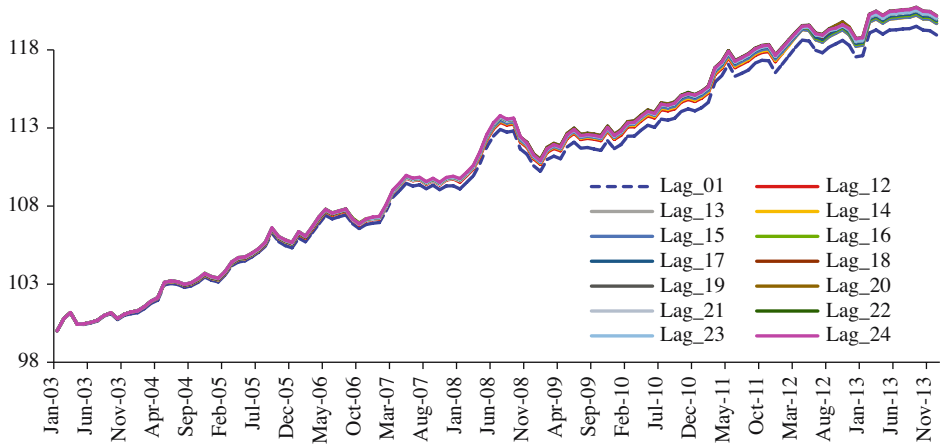

Fig. 3. Different CPI series corresponding to various implementations lags.

However, the commodity-substitution bias could generally be reduced if the implementation lag were substantially shortened. This can be shown from the difference in the growth rates between implementation lags of one month and twelve months, as well as twelve months and 24 months. Table 6 indicates that the substitution bias can be reduced from 0.221 percentage points to 0.176 percentage points if the implementation lag is shortened from 24 months to twelve months. ABS (2016) found a similar impact of the weight implementation lag on the CPI. The bias declined from 0.15% per year with a two-year implementation lag to 0.09% per year with a one-year implementation lag for the period from September 2005 to September 2011.

Table 6. Comparison of the geometric average growth rates of the different CPI series using various implementation lags and the Fisher index.

	Indexes (2003 = 100)	Differences in indexes	Annual growth rate	Difference in growth rate
	2003–2011		(%)	(%)
Fisher	114.389	0.000	1.695	0.000
Lowe index, 1 month lag	115.484	1.095	1.816	0.121
Lowe index, 12 month lag	115.980	1.591	1.870	0.176
Lowe index, 13 month lag	116.153	1.764	1.889	0.195
Lowe index, 14 month lag	116.075	1.686	1.881	0.186
Lowe index, 15 month lag	116.164	1.775	1.891	0.196
Lowe index, 16 month lag	116.300	1.911	1.905	0.211
Lowe index, 17 month lag	116.282	1.893	1.903	0.209
Lowe index, 18 month lag	116.340	1.951	1.910	0.215
Lowe index, 19 month lag	116.432	2.043	1.920	0.225
Lowe index, 20 month lag	116.413	2.023	1.918	0.223
Lowe index, 21 month lag	116.348	1.959	1.911	0.216
Lowe index, 22 month lag	116.405	2.016	1.917	0.222
Lowe index, 23 month lag	116.316	1.926	1.907	0.213
Lowe index, 24 month lag	116.393	2.004	1.916	0.221

Table 7. Different link months for introducing the 2010 CPI basket.

Possible implementing month	Possible link month	$\frac{\sum_m p_i^{2012,4} q_i^{2010}}{\sum_m p_i^{link} q_i^{2010}}$ (A) 2010 basket	$\frac{\sum_i p_i^{2012,4} q_i^{2008}}{\sum_i p_i^{link} q_i^{2008}}$ (B) 2008 basket	Difference (A)−(B)
Jan. 2012	Dec. 2011	101.707	101.589	0.118
Feb. 2012	Jan. 2012	101.263	101.192	0.071
Mar. 2012	Feb. 2012	100.821	100.773	0.048
Apr. 2012	Mar. 2012	100.382	100.370	0.011
May 2012	Apr. 2012	100.000	100.000	0.000
June 2012	May 2012	100.027	99.974	0.053
July 2012	June 2012	100.500	100.399	0.102
Aug. 2012	July 2012	100.629	100.470	0.159
Sep. 2012	Aug. 2012	100.338	100.167	0.171
Oct. 2012	Sep. 2012	100.170	100.083	0.087

From these empirical results, we cannot infer the impact on the CPI of a given link month of a particular CPI basket. To identify and illustrate this impact, we examine the introduction of a specific CPI basket. If, for example, the 2010 basket could be possibly implemented between January 2012 and October 2012, any month from December 2011 to September 2012 could, therefore, be chosen as the link month. Using Equation (28), we can determine retrospectively which month is the optimal link month for introducing the 2010 CPI basket. Table 7 shows the comparison between April 2012 and all the other possible link months, which are within the timeline of the SHS.

We obtained positive differences in the fifth column of Table 7, implying that using months either earlier or later than April 2012 as the link month cannot reduce commodity-substitution bias in the CPI based on Equation (28). Although using April 2012 as the link month to introduce the 2010 basket generates the lowest index level, it might not necessarily be true for introducing other new baskets. We therefore perform the same exercise (results are available on request) for the introduction of other baskets, and find that the optimal month for different baskets varies with the price fluctuation.

The empirical results illustrate that the impact of shortening the implementation lag on the commodity-substitution bias is not predictable, especially when the price trends are not persistent over time. However, in the case that a country's economy exhibited persistent and predictable inflation, the conditions implied by Equation (28) might very likely be satisfied. This could result in the observance of a relatively significant impact of a shortened implementation lag on the substitution bias.

Recently, as a result of operational constraints, Statistics Canada used 13 months as the implementation lag to introduce the 2011 basket. The empirical results from this study suggest that shortening the implementation lag to twelve months may not have a significant impact on further reducing the commodity-substitution bias. Moreover, the link month that yields the lowest commodity-substitution bias may not always be the same because of different monthly price fluctuations over time. As a result, it is not meaningful

to fix the link month of implementing a new basket for the purpose of reducing the commodity-substitution bias; in addition, the optimal link month of introducing a new CPI basket cannot be determined in advance. However, since Statistics Canada also compiles the CPI annual table based on the calendar year, we recommend that the new baskets be introduced in January to have a consistent annual index.

4.3. *Alternative Data Sources and Substitution Bias*

Many retailers, including nearly all major retailers, collect data through automated point-of-sale scanners. Scanner data is becoming an increasingly important source of information for statistical agencies, providing them with the prices and quantities of a large number of actual transactions in a timely manner. Several national statistical agencies currently make use of this data, including the Netherlands, Norway, Sweden, Switzerland, and New Zealand. Meanwhile, with the development of electronic commerce, online shopping has become more popular. Accompanying this growth, public information on product prices and characteristics is also available online. Automated data collection ("web-scraping") can replace traditional price collection for some product categories. With these "big data" sources, statistical agencies and academic researchers have an opportunity to study many research issues that used to be operationally infeasible and purely theoretical, and explore new methods to solve these issues.

With the availability of scanner data, it seems that the commodity-substitution bias issue raised in this article can more easily be addressed by using the prices and quantities to construct weighted (preferably superlative) price indexes. Research using scanner data to either estimate the substitution bias or to produce a superior estimate of the CPI has been ongoing for more than thirty years. New challenges and problems also arise with the use of scanner data, such as more volatile estimates of the CPI and chain-drift caused by the use of high-frequency scanner data. To overcome these new problems with the use of scanner data, Ivancic et al. (2011) proposed an innovative method, described as a rolling year GEKS method (RYGEKS), which adapts multilateral index number theory to making comparisons between multiple time periods. The GEKS method, described by Gini (1931), Eltetö and Köves (1964), and Szulc (1964), was originally used to conduct multilateral comparison, involving two stages of aggregation. The RYGEKS method makes maximum use of all matches in the scanner data to compile non-revisable CPIs that are approximately free from chain drift. Since then, this novel approach has been tested by many countries' numerical experiments, such as de Haan and van der Grient (2011) using Dutch data, Johansen and Nygaard (2011) using Norwegian data, and Krsinich (2011) using scanner data from New Zealand. Extensions to the RYGEKS have also been made; for instance, de Haan and Krisinich (2014) used an imputation Törnqvist rolling year GEKS procedure (ITRYGEKS) to derive quality-adjusted and chain-drift free price indexes. This method was applied by Statistics New Zealand (2014) to produce a CPI for its electronics products beginning in the September 2014 quarter.

Following the advancement of these new data processing techniques, the availability of large-volume data sources could provide statistical agencies with new practical solutions to primary questions covered in this paper. It might be feasible to obtain timely information on quantities purchased by households and to dramatically shorten the basket

implementation lag with the arrival of new data processing systems that could eliminate or supress some of the current operational constraints. However, more research is needed before incorporating these data sources into the CPI.

5. Conclusion

The Lowe index number formula, one of the fixed-basket concept indexes, is widely used by statistical agencies to compile their Consumer Price Index (CPI). However, because of its limitations associated with the fixed-basket concept, some concern arises from the use of this formula, in particular the issue of commodity-substitution bias. Because of the importance of the CPI to its different users (such as central banks, policy makers, and the general population as a whole), researchers have devoted, and continue to devote, much work into investigating the issue of commodity-substitution bias in the CPI.

In this article, we constructed a comprehensive data set by using information taken from Statistics Canada's Survey of Household Spending (SHS) for the years from 2000 to 2011, and the monthly CPI data for Canada at the basic class level for the period from January 2000 to December 2013.

This study focused on the investigation of approaches to reducing the commodity-substitution bias in the Canadian CPI based on two key aspects associated with the introduction of new CPI baskets. Namely, updating the CPI basket more frequently, and introducing a new CPI basket in a more timely manner. The empirical results found in this paper for the examination period indicate that increasing the frequency of updating the CPI basket could reduce the commodity-substitution bias. This finding is consistent with what has been shown in Greenlees and Williams (2009) and in other studies. In addition this paper's results reveal that the marginal gains from moving from basket-updates every four years to every two years are more significant than those from moving from basket-updates every two years to every year.

The impact of shortening the implementation lag for a new CPI basket on the commodity-substitution bias is unpredictable because it depends on the consistency between the long-term price trends (between the two basket reference periods) and the short-term price movements (between the possible link months), as well as the existence of consumers' price-induced commodity-substitution behaviour. Clear differences can be perceived in the price indexes compiled by using a twelve-month implementation lag versus an 18-month or longer implementation lag, while the differences in the indexes are largely reduced when they are constructed using twelve months compared to 14 months as implementation lags. Therefore, based on both the decomposition of index differences and the empirical results in this article, it is believed that the conclusion in Généreux (1983) would hold only when the conditions illustrated above were satisfied. Consequently, it is worthwhile for a statistical agency to pursue ways to dramatically shorten the implementation lag; however, taking great effort to slightly improve the timeliness of implementing a new basket may not provide meaningful returns.

In this article, we presented the empirical results using Canadian data for the period between 2003 and 2011. These results do not provide direct answers for choosing the most effective approach to reducing the commodity-substitution bias in a CPI. Statistical agencies in other countries can draw inferences from this empirical work but should be

cautious in generalizing these results to other CPIs because of the time dependence of the empirical results. Finally, new practical solutions associated with the incorporation of large-volume data sources in the CPI is worth further investigation from statistical agencies.

6. References

Australian Bureau of Statistics. 2016. "Increasing the Frequency of Consumer Price Index Expenditures Class Weight Updates." *Australian Bureau of Statistics Information Paper* 6401.0.60.002. July.

Bérubé, C. 1996. "Selecting a Formula for the Canadian CPI, 1962–1994." Price Division Analytical Series, Statistics Canada. No. 7, Catalogue No. 62F0014MPB, p 7.

Boskin, Michael J., Ellen R. Dullberger, Robert J. Gordon, Zvi Griliches, and Dale W. Jorgenson. 1996. "Toward a More Accurate Measure of the Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index." Washington, D.C.: US Government Printing Office.

Crawford, A. 1998. "Measurement Biases in the Canadian CPI: An Update." *Bank of Canada Review*. Spring. Page 39–56.

De Haan, Jan and Heymerik A. van der Grient. 2011. "Eliminating Chain Drift in Price Indexes Based on Scanner Data." *Journal of Econometrics* 161(1): 36–46.

De Haan, Jan and Frances Krisnich. 2014. "Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes." *Journal of Business & Economic Statistics* 32(3): 341–358.

Diewert, W. Erwin. 1976. "Exact and Superlative Index Numbers." *Journal of Econometrics.* May. Page 115–145.

Eltetö, Ö. and P. Köves. 1964. "On a Problem of Index Number Computation Relating to International Comparisons." *Statisztikai Szemle* 42: 507–518. (in Hungarian).

Généreux, Pierre A. 1983. "Impact of the Choice of Formulae on the Canadian Consumer Price Index." *Price Level Measurement: Proceedings from a Conference* Sponsored by Statistics Canada. Erwin Diewert and Claude Montmarquette (eds.).

Gini, C. 1931. "On the Circular Test of Index Numbers." *Metron* 9(9): 3–24.

Greenlees, John S. and Elliot Williams. 2009. *Reconsideration of Weighting and Updating Procedure in the US CPI*. Paper presented at the 11th Meeting of the International Working Group on Price Indexes (Ottawa group) in Neuchâtel, Switzerland.

Ho, Ricky, Peter Campion, and Chris Pike. 2011. *New Zealand Consumer Price Index—an empirical analysis of the frequency and level of weight updates*. Room document at the 12th Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand.

ILO / IMF / OECD / IMECE / Eurostat / World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO.

Ivancic, Lorraine, W. Erwin Diewert, and Kevin J. Fox. 2011. "Scanner Data, Time Aggregation and Construction of Price Indexes." *Journal of Econometrics* 161(1): 24–35.

Johansen, Ingvild, and Ragnhild Nygaard. 2011. *Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages*. Paper presented at the 12th

Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand.

Krsinich, Frances. 2011. *Price indexes from scanner data*. Paper presented at the 12th Meeting of the International Working Group on Price Indexes (Ottawa group) in Wellington, New Zealand. May 2011.

Rossiter, James. 2005. *Measurement bias in the Canadian Consumer Price Index*. Bank of Canada Working Paper 2005-39. December.

Sabourin, Patrick. 2012. "Measurement bias in the Canadian Consumer Price Index: An update." *Bank of Canada Review*. Summer.

Shiratsuka, Shigenori. 2006. "Measurement Errors in Japanese Consumer Price Index." *Monetary and Economic Studies* 17(3). Institute for Monetary and Economic Studies, Bank of Japan, 69–102.

Statistics Canada. 2014. "The Canadian Consumer Price Index Reference Paper." Catalogue No. 62-553-X.

Statistics New Zealand. 2014. *Measuring price change for consumer electronics using scanner data*. Available from www.stats.govt.nz.

Szulc, B. 1964. "Indexes for Multiregional Comparisons." *Przeglad Statystyczny* 3: 239–254. (in Polish).

Wynne, Mark A. and Diego Rodriguez-Palenzuela. 2002. *Measurement Bias in the HICP: What do we know, and what do we need to know?* European Central Bank Working Paper Series, No. 131. Frankfurt: European Central Bank.

Received January 2016
Revised October 2016
Accepted January 2017

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1021–1050, http://dx.doi.org/10.1515/JOS-2017-0048

Estimating Cross-Classified Population Counts of Multidimensional Tables: An Application to Regional Australia to Obtain Pseudo-Census Counts

Thomas Suesse[1], *Mohammad-Reza Namazi-Rad*[1], *Payam Mokhtarian*[3], *and Johan Barthélemy*[2]

Estimating population counts for multidimensional tables based on a representative sample subject to known marginal population counts is not only important in survey sampling but is also an integral part of standard methods for simulating area-specific synthetic populations. In this article several estimation methods are reviewed, with particular focus on the iterative proportional fitting procedure and the maximum likelihood method. The performance of these methods is investigated in a simulation study for multidimensional tables, as previous studies are limited to 2 by 2 tables. The data are generated under random sampling but also under misspecification models, for which sample and target populations differ systematically. The empirical results show that simple adjustments can lead to more efficient estimators, but generally, at the expense of increased bias. The adjustments also generally improve coverage of the confidence intervals. The methods discussed in this article along with standard error estimators, are made freely available in the R package `mipfp`. As an illustration, the methods are applied to the 2011 Australian census data available for the Illawarra Region in order to obtain estimates for the desired three-way table for age by sex by family type with known marginal tables for age by sex and for family type.

Key words: Census data; IPFP; Log-linear model; model-based inference; count estimation; synthetic population.

1. Introduction

In many countries, census data are still the major source for geographically detailed estimates of populations and economies. Statistical agencies often provide public-use microdata files based on their census or surveys. To preserve confidentiality, some variables might be suppressed, or alternatively only marginal totals, also known as aggregated data, are released instead of the joint totals. For example, joint tables on age by sex and by income might not be released for small areas, as this could lead to disclosing the income of some people with specific age by sex. Instead, only separate tables of marginal totals, for example for age by sex and for income, are released.

[1] National Institute for Applied Statistics Research Australia, School of Mathametics and Applied Statistics, University of Wollongong, NSW 2522, Australia. Email: tsuesse@uow.edu.au
[2] SMART Infrastructure Facility, University of Wollongong, NSW 2522, Australia
[3] Damian Group, Fairfax Media, Sydney 2009 NSW, Australia

Acknowledgments: The authors wish to gratefully acknowledge the help of Dr Madeleine Strong Cincotta in the final language editing of this article. We are grateful to the associate editor and the anonymous referees for their helpful comments that greatly improved the article.

The generation of an artificial (or synthetic) population that realistically matches the population of interest from such limited tables, or more generally aggregated data, has become an important research area (Arentze et al. 2007; Gargiulo et al. 2010; Harland et al. 2012; Barthélemy and Toint 2013; Lenormand and Deffuant 2013; Geard et al. 2013; Huynh et al. 2016).

The conventional generation process of such a synthetic population (SP) is a two step procedure that integrates data from a fully disaggregated sample (for example derived from a survey) with aggregated data from a census (Beckman et al. 1996). The first step estimates the contingency table of all the attributes for the area of interest. The second step then randomly draws synthetic individuals from the sample in proportions that match the estimated contingency table. Using this SP generation approach, the risk of identification of population units and/or their sensitive information in the generated synthetic data is greatly reduced (Rubin 1987).

This article focuses on the first step, that is, the estimation of population counts in multidimensional contingency tables when a random sample is available together with known marginal population tables of lower dimension. It is also important to investigate the multidimensional case with several variables in which each variable has possibly more than two categories, as existing simulation studies, for example by Little and Wu (1991), only considered the unrealistic scenario of a 2 by 2 table.

The iterative proportional fitting procedure (IPFP) originally proposed by Deming and Stephan (1940) and the maximum likelihood (ML) method (Smith 1947) are the traditional methods for estimating cross-classified population counts. IPFP is a general purpose method to match marginal information and is not limited to surveys. The method has also been applied in small area estimation (SAE) to a slightly different situation when the complete table is replaced by some other source of information, such as a complete table from a previous census together with marginal tables which are not necessarily known but are based on some survey estimates. In this context, the method is known as structure preserving estimation (SPREE) (Purcell and Kish 1980; Zhang and Chambers 2004), as it preserves part of the structure of the implied log-linear models in both tables. IPFP has the same structure preserving property and SPREE can be thought of as a special case of IPFP. For example, Purcell and Kish (1980) have considered six different data situations and only referred to one as IPFP; however, all six situations were indeed solved with IPFP.

Section 2 introduces the main estimation methods IPFP and ML, as well as two other estimation methods. Data adjustments are also introduced, which are applied before the estimators are calculated in order to improve statistical properties. In Section 3, misspecification models are considered, that is models for which sample and population information differ systematically, including ML estimators for each of these misspecification models. In Section 4, a simulation-based empirical study is presented to investigate the performance of the methods discussed in this article under simple random sampling and under the misspecification models. The methods are then employed for estimating cross-classified population counts and probabilities for the Illawarra region using available one- and two-dimensional 2011 Australian census tables. This article concludes with a discussion of the results.

2. Estimating Cross-Tabulated Population Counts

The main methods for estimating cross-tabulated population counts and probabilities subject to known marginal population tables of lower dimensions are discussed in this section.

2.1. Iterative Proportional Fitting Procedure (IPFP)

IPFP was originally proposed by Deming and Stephan (1940) as an algorithm attempting to minimize the Pearson chi-squared statistic. For the purpose of population reconstruction, IPFP is often used as an algorithm attempting to adjust census tables so that table cells add up to totals in all required dimensions (Fienberg 1970; Gargiulo et al. 2010; Farooq et al. 2013; Barthélemy and Toint 2013). This application of iterative proportional fitting (IPF) to contingency tables with known margins is called raking (Stephan 1942). Raking (also known as raking ratio estimation) is a procedure which applies a proportional adjustment to the sample weights in a survey so that the adjusted weights add up the known population total when only the marginal population totals are known (Deville et al. 1991; Lu and Gelman 2003). Although raking is not a maximum likelihood (ML) method under random sampling, the raking estimates are consistent and best asymptotically normal (Arentze et al. 2007).

Ireland and Kullback (1968) showed that the estimator produced by the IPFP method minimizes the discrimination information criterion (also known as the Kullback-Leibler divergence, or relative entropy). Mosteller (1968) pointed out that IPFP also preserves the interaction structure of the initial table as defined by the conditional odd ratios.

For illustration purposes, we restrict ourselves to three-way tables, but the methods can be applied in a straightforward manner to tables with more variables. For a three-way contingency table referring to three categorical variables X_1, X_2, and X_3 each with A, B, and C levels, respectively, the population counts are denoted by N_{abc} with population size $N = \sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} N_{abc} = N_{\bullet\bullet\bullet}$, where the dot (i.e., $\bullet$) refers to summation over the corresponding variable. The one-way marginal cell counts $N_{a\bullet\bullet}, N_{\bullet b\bullet}$, and $N_{\bullet\bullet c}$ are defined accordingly, for example $N_{a\bullet\bullet} = \sum_{b=1}^{B}\sum_{c=1}^{C} N_{abc}$. The two-way marginal totals are denoted by $N_{ab\bullet}$, $N_{a\bullet c}$, and $N_{\bullet bc}$ and defined by summing the N_{abc} over the respective index.

The main objective is to estimate the cell probabilities $\pi_{abc} = P(X_1 = a, X_2 = b, X_3 = c)$, or equivalently N_{abc}. All joint probabilities π_{abc} and marginal probabilities, such as $\pi_{ab\bullet}$ and $\pi_{a\bullet\bullet}$, need to sum up to one, as marginal probabilities also characterize a valid discrete distribution. When dealing with sample data, sample counts are denoted by y_{abc} with $n = y_{\bullet\bullet\bullet}$ denoting the total sample size.

In the classical IPFP presented by Deming and Stephan (1940), the initial value for the cell probabilities are set as $\pi_{abc}^{(0)} = (ABC)^{-1}$, which corresponds to the case of having no sample data available. When using IPFP for population synthesis, the initial cell probabilities are based on representative survey data with counts y_{abc} often referred to as the *seed data*, that is $\pi_{abc}^{(0)} = y_{abc}/n$. Let us assume for illustration purposes that the three two-way marginal population counts $N_{ab\bullet}$, $N_{a\bullet c}$, and $N_{\bullet bc}$ are available. We aim at finding

π_{abc} so that the following population constraints hold

$$\pi_{ab\bullet} = \frac{N_{ab\bullet}}{N}, \pi_{a\bullet c} = \frac{N_{a\bullet c}}{N} \text{ and } \pi_{\bullet bc} = \frac{N_{\bullet bc}}{N}. \quad (1)$$

Then one iteration of the IPFP consisting of a three-step cycle has the form

$$\pi_{abc}^{(k+1)} = \frac{\pi_{abc}^{(k)}}{\sum_{a=1}^{A} \pi_{abc}^{(k)}} \times \pi_{\bullet bc}, \quad \pi_{abc}^{(k+2)} = \frac{\pi_{abc}^{(k+1)}}{\sum_{b=1}^{B} \pi_{abc}^{(k+1)}} \times \pi_{a\bullet c},$$

$$\pi_{abc}^{(k+3)} = \frac{\pi_{abc}^{(k+2)}}{\sum_{c=1}^{C} \pi_{abc}^{(k+2)}} \times \pi_{ab\bullet}.$$

The algorithm is continued by setting $k := k + 3$ until convergence to the desired accuracy is attained. Importantly, the obtained estimates $\hat{\pi}_{abc} = \pi_{abc}^{(k)}$ will satisfy (1). The algorithm will converge to a unique solution provided the seed data contain strictly positive entries and provided the marginal constraints do not contradict each other. For example, the constraints $N_{ab\bullet}$ and $N_{a\bullet c}$ need to result in the same $N_{a\bullet\bullet}$, that is $N_{a\bullet\bullet} = \sum_b N_{ab\bullet} = \sum_c N_{a\bullet c}$.

Setting positive starting values $\left(\pi_{abc}^{(0)} > 0\right)$ ensures that each cell has a non-zero probability estimate, that is $\hat{\pi}_{abc} > 0$ (Gange 1995). If some zero cell counts are observed, that is $y_{abc} = 0$, then adjustments can be made. For example adding the value of 0.5 to all cells, the standard procedure for 2 by 2 tables (Agresti 2002, 71). An alternative proposed by Lang (2004) is to add a tiny constant (e.g., 10^{-6}) to all the zero cells to ensure that the estimates are strictly positive, that is $\hat{\pi}_{abc} > 0$.

Let $\boldsymbol{\pi}$ denote the vector $\boldsymbol{\pi} = (\pi_{111}, \ldots, \pi_{11C}, \ldots, \pi_{AB1}, \ldots, \pi_{ABC})^T$ of length $K = ABC$ and let the $AB + CB + AC$ constraints $N_{ab\bullet}/N$, $N_{a\bullet c}/N$ and $N_{\bullet bc}/N$ be stored in vector $\mathbf{c}$ and let matrix $\mathbf{A}$ be the $(AB + CB + AC) \times K$ matrix such that $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$. Then, following Little and Wu (1991), a (co)variance estimator for $\hat{\boldsymbol{\pi}}$ is:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}^{-1}(\mathbf{p})\mathbf{U})(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}\mathbf{U}^T, \quad (2)$$

where $\mathbf{D}(\mathbf{a})$ is the diagonal matrix having vector $\mathbf{a}$ on its diagonal, and $\mathbf{p}$ is the vector of sample proportions, that is $\mathbf{p} = (p_{111}, \ldots, p_{11C}, \ldots, p_{AB1}, \ldots, p_{ABC})^T$ with $p_{abc} = y_{abc}/n$. Matrix $\mathbf{U}$ is an orthogonal complement of $\mathbf{A}$, such that $\mathbf{A}^T\mathbf{U} = 0$ and $(\mathbf{A}, \mathbf{U})$ has full rank. To achieve the full rank matrix $(\mathbf{A}, \mathbf{U})$, matrix $\mathbf{A}$ also needs to be of full rank. This requires removing three elements in vector $\mathbf{c}$ (and the corresponding rows in $\mathbf{A}$), as the second order constraints are linearly dependent, for example $N_{AB\bullet} = N - \sum_{a=1}^{A-1}\sum_{b=1}^{B-1} N_{ab\bullet}$.

Even though IPFP is often used to obtain population estimates $\hat{N}_{abc}$ via the simple formula

$$\hat{N}_{abc} = N\hat{\pi}_{abc}, \quad (3)$$

the (co)variance formula, see (2), to obtain confidence intervals for these population estimates is often not discussed in the literature on SP generation and is worth highlighting, as it provides an uncertainty measure.

It should be noted that the raking estimates denoted by $\hat{\pi}^r_{abc}$ based on all three second order population constraints are of the following form (Little and Wu 1991)

$$\log\left(\frac{\hat{\pi}^r_{abc}}{p_{abc}}\right) = \hat{\theta}^r + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)} + \hat{\theta}^r_{12(ab)} + \hat{\theta}^r_{13(ac)} + \hat{\theta}^r_{23(bc)}, \tag{4}$$

where θ are suitable parameters.

2.2. *Maximum Likelihood Approach*

The maximum likelihood method under random sampling (MLRS) has been considered for 2 by 2 tables by Smith (1947) and Little and Wu (1991) but has not been particularly extensively studied when dealing with more than two variables. For a three-way contingency table, Equation (1) can be expressed as $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$ with linearly dependent constraints removed.

Let $\mu_{abc} = E(y_{abc})$ with the corresponding vector $\boldsymbol{\mu}$ defined in a similar fashion as $\boldsymbol{\pi}$ and define the function $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\pi} - \mathbf{c}$ with $\boldsymbol{\pi} = \boldsymbol{\mu}/n$. With this definition, $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$ when $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$ and $\mathbf{h}(\boldsymbol{\mu}) \neq 0$ otherwise. Lang and Agresti (1994) and Lang (1996, 2004, 2005) provide a model framework and base estimation on maximising the log-likelihood subject to some arbitrary constraints expressed by $\mathbf{h}(\boldsymbol{\mu}) = 0$. This is achieved by using the famous method of Lagrange multipliers, which maximizes the constrained log-likelihood L_c

$$L_c = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}), \tag{5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{AB-1}, \ldots, \lambda_{AB+BC+AC-3})^T$ is a vector of the so-called Lagrange multipliers.

Joseph B. Lang provides an `R` function (mph.fit) available on http://homepage.stat.uiowa.edu/~jblang/mph.fitting/ for ML estimation of multinomial-Poisson homogeneous (MPH) models for contingency tables. Bergsma et al. (2009) provide a more efficient algorithm (`R` package `cmm`) to fit such models. Apart from obtaining estimates $\hat{\boldsymbol{\mu}}$, that will satisfy the population constraints, the ML method also provides a (co)variance matrix for $\hat{\boldsymbol{\mu}}$ as follows:

$$\widehat{\text{Cov}}\,(\hat{\boldsymbol{\mu}}) = \mathbf{D}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T/n - \mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H}(\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}}), \tag{6}$$

where $\mathbf{H}(\boldsymbol{\mu}) = \frac{\partial \mathbf{h}^T(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ (Lang 2004).

Compared to log-linear models of the form $\log(\boldsymbol{\mu}) = \mathbf{X}\beta$ with design matrix $\mathbf{X}$ and the vector of fixed effects parameters $\boldsymbol{\beta}$, see, for example Agresti (2002), Formula (6) shows an additional term (the last term). This additional term reduces the variance imposed by the restrictions or constraints compared to the unconstrained model. Little and Wu (1991) proposed a different (co)variance formula for the ML method based on the delta method similar to (2) and is given by:

$$\widehat{\text{Cov}}\,(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}\mathbf{U}^T. \tag{7}$$

To obtain model-based population estimates $\hat{N}_{abc}$, Formula (3) is applied. Finally, the estimated (co)variance of the estimated population counts contained in the vector $\hat{\mathbf{N}} = \frac{N}{n}\hat{\boldsymbol{\mu}}$ is:

$$\widehat{\text{Cov}}\,(\hat{\mathbf{N}}) = N^2\widehat{\text{Cov}}\,(\hat{\boldsymbol{\pi}}). \tag{8}$$

It should be noted that the ML estimates $\hat{\pi}_{abc}^{ML}$ (based on second order population constraints) are of the form (see Appendix A)

$$\left(\frac{\hat{\pi}_{abc}^{ML}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}_{1(a)}^{ML} + \hat{\theta}_{2(b)}^{ML} + \hat{\theta}_{3(c)}^{ML} + \hat{\theta}_{12(ab)}^{ML} + \hat{\theta}_{13(ac)}^{ML} + \hat{\theta}_{23(bc)}^{ML}. \tag{9}$$

This ML method can also be used to fit standard log-linear models by including the model in the constraint function $\mathbf{h}(\boldsymbol{\mu})$ by setting $\mathbf{h}(\boldsymbol{\mu}) = \tilde{\mathbf{U}}^T \log \boldsymbol{\mu} = 0$, where $\tilde{\mathbf{U}}$ is a full column rank orthogonal complement of the design matrix $\mathbf{X}$. Lang (1996, 2004, 2005) extended this methodology to generalized log-linear models and homogeneous linear predictor models.

2.3. Other Estimation Methods

Two other popular estimation methods are the least squares method (LSQ) and the minimum chi-squared method (CHI2), see Little and Wu (1991). The LSQ estimates are obtained by minimizing the following criteria

$$\sum_{abc} \frac{(y_{abc} - n\pi_{abc})^2}{y_{abc}} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}), \tag{10}$$

and CHI2 estimates by minimizing

$$\sum_{abc} \frac{(y_{abc} - n\pi_{abc})^2}{\pi_{abc}} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}). \tag{11}$$

Similar to the ML method, solutions to (10) and (11) can be obtained by applying the Lagrange multiplier method. Under simple random sampling these methods are not recommended, because ML and IPFP provide more efficient estimates (Little and Wu 1991).

2.4. Adjusted Estimators

We also consider adjusted estimation methods "$+\alpha$", where the value of α is added to all counts y_{abc} before the estimators are calculated. Two such typical adjustments for contingency tables, namely $\alpha = 0.5$ and α being a tiny constant (the former to all counts and the latter only to zero counts), are mentioned in Subsection 2.1. Such adjustments of the data could lead to more efficient estimators. To further justify these adjustments, consider that the standard ML estimator for the multinomial distribution with K (here $K = ABC$) probabilities π_i and size n is $p_i = y_i/n$ for $i = 1, \ldots, K$ (the sample proportions). Under the Bayesian approach, the so-called Dirichlet distribution with parameters α_i is the conjugate family of priors for the multinomial distribution. The posterior distribution is also a Dirichlet distribution with parameters $y_i + \alpha_i$. The Bayesian estimate for π_i (the posterior mean) is $\frac{y_i+\alpha_i}{n+M}$ with $M = \sum_i \alpha_i$. Let $\gamma_i = \alpha_i/M$, then the Bayesian estimator (posterior mean) equals the weighted average

$$\frac{n}{n+M} \times p_i + \frac{M}{n+M} \times \gamma_i$$

with weights summing to one, that is $\frac{n}{n+M} + \frac{M}{n+M} = 1$ (Agresti and Hitchcock 2005).

A standard non-informative prior, the uniform prior, is obtained by setting $\alpha_i = 1$ (Jeffreys 1998; Gelman et al. 2003). In the binomial case ($K = 2$), this corresponds to assuming a uniform prior for π_i and leads to the Bayesian estimator $\frac{y_i+1}{n+2}$. In the multinomial case, this leads to $\frac{y_i+1}{n+K}$. In general, when adding α_i to cell y_i yielding new counts $\tilde{y}_i = y_i + \alpha_i$ and then applying the ML estimator $\tilde{y}_i/\tilde{n}$ with new sample size $\tilde{n} = n + M$ yields the Bayesian estimator, because $\frac{\tilde{y}_i}{\tilde{n}} \equiv \frac{y_i+\alpha_i}{n+M}$.

The case $\alpha_i = 1/2$ leads to the popular Jeffrey's prior. Often any prior with $\alpha_i = \alpha$ may be considered as non-informative (De Campos and Benavoli 2011). In this sense, "+0.5" and "+1" are special cases of non-informative priors. However, the concept of non-informative priors is debated in the Bayesian community, for example the uniform prior can be considered as highly informative and the Jeffrey's prior as non-informative, or vice versa. Agresti and Hitchcock (2005) noted the "lack of consensus about what 'non-informative' means". In this article, we consider all choices of a constant $\alpha_i = \alpha$ as non-informative following De Campos and Benavoli (2011).

The classical ML and Bayesian estimators apply to the unrestricted case (without imposing marginal constraints). When the margins/constraints are met by these classical estimators, however, they coincide with the restricted estimators, as $\boldsymbol{\lambda} = 0$ in this case, see Equation (5). If the prior distribution reflects the true sampling mechanism, then the Bayesian estimator has the highest efficiency by construction, as it is widely known that the (posterior) mean minimizes the mean squared error. Assuming the π_{abc} are not known and can be of any size, we anticipate that a choice of a "non-informative prior" ($\alpha > 0$) could also lead to improved efficiency in the restricted case.

Adjusting the observations and then applying standard Wald type confidence intervals (CI) has been applied by Agresti and Coull (1998), where the number of failures and successes was increased by two before the Wald type CI was applied (method "+2"). This method – the so-called Agresti-Coull-Interval – yielded better coverage than the standard Wald-type CI without "data adjustment" and even better coverage than the "exact" CI, in the sense that better means closer to the nominal 95% level, as the "exact" method is highly conservative. Based on these different choices for "+α" in the literature we also consider the methods "+2" and "+10" to investigate the effect of choosing a different value of $\alpha_i = \alpha$.

3. Misspecification Models

In theory, a probability sample is taken from a population, implying that both sample and population have the same characteristics. In practice, however, samples can differ systematically from the target population, due to, for example, omission of units or errors in the sampling frame, or very commonly due to the nonresponse of some selected units.

Let us now assume that the sample was obtained from a population, now referred to as the non-target population, which is not the same as the target population, the population of interest. We denote the probabilities referring to the non-target population by τ_{abc} and those to the target population by π_{abc}. Following Little and Wu (1991), we consider the following models relating π_{abc} and τ_{abc}

$$\left(\frac{\pi_{abc}}{\tau_{abc}}\right)^{\kappa} = \theta_\kappa + \theta_{1(a)\kappa} + \theta_{2(b)\kappa} + \theta_{3(c)\kappa} + \theta_{12(ab)\kappa} + \theta_{13(ac)\kappa} + \theta_{23(bc)\kappa}, \qquad (12)$$

where $\kappa = -1, 1, 2$ and $\kappa \to 0$ (in the following denoted by $\kappa = 0$) refers to the log-function, that is $\log\left(\frac{\pi_{abc}}{\tau_{abc}}\right)$. The specification of the θ parameters in (12) implies that second order population margins are provided. These four models specified by the value of κ provide flexible adjustments when sample and target population characteristics do not agree. Following similar arguments for the two-dimensional case as in Little and Wu (1991), we can show that the ML estimates for the model $\kappa = 0$ are provided by IPFP, see Appendix B. Similarly, it can be shown that the ML estimators for $\kappa = 1, -1, -2$ are of specific form. To summarize, the following results hold (Little and Wu 1991):

- For $\kappa = 1$: ML estimates are provided by LSQ
- For $\kappa = 0$: ML estimates are provided by IPFP
- For $\kappa = -1$: ML estimates are provided by ML(RS)
- For $\kappa = -2$: ML estimates are provided by CHI2.

The proofs for $\kappa = 1, -1, -2$ in the three-dimensional case are not shown to preserve space, but follow similar arguments as for $\kappa = 0$ and the two dimensional case.

Little and Wu (1991) compared all four estimation methods (IPFP, ML, CHI2, LSQ) in a simulation study while simulating data using random sampling and under each of the four misspecification models. The averaged results over a wide range of settings (Little and Wu 1991, see Table 1) show that under all five situations (random sampling and the four misspecification models), either IPFP or MLRS are the best performing methods. MLRS is best under random sampling and for the models $\kappa = -1, 2$, whereas IPFP is best under the models $\kappa = 0$ and $\kappa = 1$. To be more precise, MLRS is best for $\kappa = -1$ and IPFP is best for $\kappa = 0$, as expected but LSQ and IPFP are well performing methods for $\kappa = 1$ with IPFP being slightly better. Similarly, MLRS and CHI2 perform best when $\kappa = -2$, with a slightly better performance of MLRS than CHI2.

Even though these results are averaged over all simulations and limited to 2 by 2 tables, they still show that the commonly used IPFP and MLRS methods generally perform well, but their results refer to a single 2 by 2 table, an unrealistic situation for often sparse multidimensional tables. The next section considers a simulation study specially designed for multidimensional tables with a large number of cells.

4. Simulation Study

4.1. Setup

Little and Wu (1991) and Causey (1983) conducted empirical simulation studies based on 2 by 2 tables with constraints referring to the two (marginal) variables. It is not clear how these results can be extended to multidimensional tables with a large number of cells and more than two sets of population constraints.

When obtaining a sample (table) with small n from a population (table) with many cells, the sample table is often sparse. The simulation study considers table cells with low to relatively large probabilities by setting $A = 5$, $B = 4$ and $C = 2$ and where the $K = ABC = 40$ probabilities $\boldsymbol{\pi} = (\tilde{\pi}_{111}, \ldots, \tilde{\pi}_{11c}, \ldots, \tilde{\pi}_{AB1}, \ldots, \tilde{\pi}_{ABC})^T = (\tilde{\pi}_1, \tilde{\pi}_2 \ldots, \tilde{\pi}_K)^T$ are monotone increasing and the kth probability is $\tilde{\pi}_k \propto \exp([5(k-1)+1]/40)$ (proportional to exponential function and then normalized to sum to one), yielding

$\tilde{\pi}_{111} = \tilde{\pi}_1 = 0.0009 < \ldots < \tilde{\pi}_{ABC} = \tilde{\pi}_{40} = 0.1183$. We consider simple random sampling (RND) and the misspecification models in Section 3.

For each of these models, we sample randomly 10,000 population tables, where each table contains randomly generated counts denoted by y_{abc}^{pop} from a multinomial distribution with parameters $\tilde{\boldsymbol{\pi}}$ and N. The aim is to estimatethe population cell probabilities denoted by $\pi_{abc} = y_{abc}^{pop}/N$. For small near zero $\tilde{\pi}_{abc}$, the obtained y_{abc}^{pop} are often zero. This is a realistic scenario for multidimensional tables, as in practice some population counts will indeed be small and often be zero.

In Section 5, we use individual level sample data from a larger area (n large) to estimate population totals of a smaller area (N), such as $n > N$. This scenario does not warrant random sampling without replacement (for which $n < N$) and requires the consideration of misspecification models.

For simplicity, the misspecification models specified by (12) only include main effects $\theta_{1(a)\kappa}, \theta_{2(b)\kappa}, \theta_{3(c)\kappa}$, which are generated under a $N(\mu = 0, \sigma^2 = 1)$ distribution for $\kappa = 0$ and from a lognormal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ for $\kappa = 1, -1, -2$ to have strictly positive parameters in the latter case. Rearranging Equation (12) in terms of τ_{abc} for $\kappa = 1, -1, -2$ gives

$$\tau_{abc} = \pi_{abc} \times (\eta_{abc})^{-\frac{1}{\kappa}}$$

and for $\kappa = 0$

$$\tau_{abc} = \pi_{abc} \times \exp(-\eta_{abc}),$$

where $\eta_{abc} = \theta_\kappa + \theta_{1(a)\kappa} + \theta_{2(b)\kappa} + \theta_{3(c)\kappa}$. The constant θ_κ is chosen such that all τ_{abc} sum to one. Then based on these τ_{abc}, a sample of size n can be obtained by random sampling to estimate the π_{abc}.

We investigate the performance of the estimators IPFP, MLRS (abbreviated here ML), LSQ, and CHI2 and their (co)variance estimators, and their adjusted versions "+0", "+0.5", "+1", "+2", and "+10" and any combination thereof, for example IPFP + 0 and CHI2 + 10.

To assess the efficiency we calculate the mean squared error (MSE) of IPFP + 0 and the relative MSE (RMSE) of all other methods relative to IPFP + 0. As each table has many cells, the MSE for a cell is defined as $E(\hat{\pi}_{abc} - \pi_{abc})^2$. The RMSE is always relative to IPFP + 0. A value greater than one indicates a larger MSE than the MSE of IPFP + 0 and a value less than one indicates a more efficient estimator. The bias is also assessed by calculating the relative bias, here defined as $E(\hat{\pi}_{abc} - \pi_{abc})/E(\pi_{abc})$.

For the confidence intervals (CI) and the ML method we consider Lang's formula and the delta method, see Formulae (6) and (7), however, due to similar performances and a generally slightly better performance of Lang's formula, we only show results based on Lang's formula.

One of the main questions is which (model-based) estimation method is best. If a true random sample is obtained, then the ML method is the appropriate choice. The package `mipfp` provides results of several goodness of fit (GOF) tests, such as Pearson's score test X^2, the Likelihood-Ratio statistic G^2 and the Wald statistic W^2, developed by Lang (2004). They test essentially whether the sample agrees with the population; in formula $H_0 : \mathbf{h}(\boldsymbol{\mu}) = 0$ versus $H_1 : \mathbf{h}(\boldsymbol{\mu}) \neq 0$. If H_0 is rejected, then there is a strong indication that the 'sample' is not a real random sample from the target population and one of the

misspecification models should be considered. To assess whether these GOF tests are useful for determining whether a true sample is provided or rather a misspecification model applies, we also recorded the rejection rate of the GOF tests. For zero cell counts or zero estimates these might not always be calculable and adjusted versions X^2_{adj}, G^2_{adj} and W^2_{adj} are considered. Similar to Lang's mph.fit implementation, X^2_{adj} is calculated over those cells for which $\hat{\pi}_{abc} > 0$, G^2_{adj} over those for which $\hat{\pi}_{abc} > 0$ and $y_{abc} > 0$, and W^2_{adj} over those for which $y_{abc} > 0$. Without these adjustments, the statistics would be undefined for many data sets, as they would contain zero valued denominators.

To illustrate the methods on higher dimensions, we also consider the five dimensional case where each dimension has three categories leading to $K = 3^5 = 243$ probabilities with $\tilde{\pi}_k \propto \exp(k/243)$.

4.2. Results

Table 1 shows the results of the MSE/RMSE for the IPFP and ML methods, similarly Table 2 shows the relative bias and Table 3 the coverage for these two methods along with the adjusted versions. Similar tables for LSQ and CHI2 are shown in Appendix D. The tables show n, N, the model (either RND or $\kappa = 0, 1, -1, 2$), and the expected probability $E(\pi) = E(\pi_{abc}) = \tilde{\pi}_{abc}$, calculated over all tables. The values of $E(\pi)$ are chosen such that the impact of relatively small, medium-sized and large probabilities (and likewise counts) can be observed, as different methods are expected to perform differently for cells of different sizes. The values of $E(\pi)$ are ordered, such that the first row under each configuration represents the smallest $E(\pi)$ (e.g., 0.09%), the second is the first tertile, a tertile is defined as the first three-quantile, (e.g., 0.40%), the third is the median (e.g., 0.97%) and the fourth is the largest of the $E(\pi)$ (e.g., 11.83%).

For RND, we observe that all methods (IPFP, ML, CHI2 and LSQ) perform similarly but ML is still generally the best in terms of efficiency and the smaller the n is, the larger is the performance improvement. The efficiency generally improves when $\alpha > 0$. For $N = 10{,}000$ and $n = 600$, "+10" appears best, whereas for $N = 600$ and $n = 200$ the size of α that has highest efficiency depends on $E(\pi)$. For small $E(\pi)$, "+0.5" appears best, for medium and large $E(\pi)$ it is "+2" and "+10". As can be seen in Table 2, the drawback of larger α is that the bias generally deteriorates. Table 3 shows that the coverage of the unadjusted "+0" method is often too low and improved by the adjusted methods $\alpha > 0$, an optimum appears around "+0.5" and "+1". An exception appears for the five-dimensional case and large $E(\pi)$, for which $\alpha = 0.5$ appears to lower the coverage slightly, but it still increases coverage significantly for small and medium sized $E(\pi)$.

Let us now focus on the misspecification models. When evaluating the unadjusted versions in terms of efficiency, for large $E(\pi)$ the ML method under the respective misspecification model is best. For example, CHI2 is best under $\kappa = -2$ when $E(\pi) = 11.83\%$. The results of Little and Wu (1991) are confirmed in the sense that under models $\kappa = 0$ and $\kappa = 1$ the LSQ and IPFP methods perform well, and under $\kappa = -1$ and $\kappa = -2$ the ML and CHI2 methods perform well.

It also apparent, however, that for large $E(\pi)$, ML is not efficient for $\kappa = 0, 1$, whereas IPFP is a well performing method for all κ. Overall, it appears that IPFP is the best method, as it performs well regardless of the misspecification model and the size of $E(\pi)$. In terms

Table 1. $E(\pi)$ in percentages, $10^5 \times$ actual MSE for IPFP + 0 (highlighted in bold) and RMSE for other methods relative to MSE of IPFP + 0, RMSE $<$ 1 indicates better and RMSE $>$ 1 worse, all based on 10,000 simulated data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	**0.159**	0.362	0.237	0.172	0.120	1.027	0.360	0.240	0.186	0.200
0.40	**0.566**	0.666	0.485	0.302	0.085	1.003	0.663	0.480	0.294	0.076
0.97	**1.303**	0.855	0.742	0.580	0.198	0.996	0.851	0.739	0.577	0.197
11.8	**9.297**	0.974	0.963	0.955	0.937	0.996	0.970	0.960	0.957	1.157
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	**0.363**	0.290	0.302	0.333	0.368	1.092	0.292	0.311	0.355	0.431
0.40	**1.443**	0.380	0.293	0.276	0.345	1.005	0.375	0.288	0.273	0.352
0.97	**2.924**	0.654	0.500	0.378	0.352	0.993	0.647	0.494	0.373	0.368
11.8	**19.73**	0.932	0.903	0.863	0.633	0.991	0.925	0.903	0.897	1.127
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	**0.098**	0.673	0.742	0.940	1.506	1.022	0.641	0.690	0.864	1.497
0.40	**0.413**	0.759	0.668	0.642	1.048	0.982	0.722	0.630	0.592	0.945
0.97	**1.029**	0.877	0.790	0.687	0.702	0.935	0.810	0.730	0.632	0.616
11.8	**8.618**	0.978	0.971	0.976	1.141	0.732	0.717	0.714	0.718	0.937
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	**0.302**	0.559	0.597	0.634	0.643	1.308	0.530	0.463	0.445	0.525
0.40	**1.908**	0.407	0.404	0.425	0.458	1.553	0.769	0.594	0.451	0.309
0.97	**5.730**	0.460	0.428	0.425	0.432	2.416	1.563	1.287	0.989	0.446
11.8	**51.19**	0.576	0.604	0.689	0.994	4.655	3.740	3.653	3.505	2.764
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	**0.094**	0.688	0.771	0.986	1.586	1.090	0.699	0.728	0.900	1.581
0.40	**0.418**	0.755	0.671	0.652	1.037	1.181	0.845	0.722	0.646	0.934
0.97	**1.074**	0.888	0.815	0.733	0.776	1.315	1.153	1.037	0.885	0.708
11.8	**8.726**	0.989	0.992	1.012	1.229	1.724	1.688	1.663	1.629	1.620
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	**0.080**	0.709	0.778	1.016	1.795	1.024	0.703	0.768	1.004	1.883
0.40	**0.356**	0.802	0.715	0.678	1.090	0.988	0.789	0.700	0.660	1.077
0.97	**0.886**	0.908	0.839	0.749	0.740	0.983	0.892	0.824	0.734	0.716
11.8	**6.518**	0.985	0.981	0.989	1.148	0.925	0.910	0.906	0.915	1.198
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	**0.021**	0.171	0.142	0.122	0.079	0.999	0.206	0.204	0.223	0.255
0.07	**0.116**	0.174	0.132	0.113	0.083	0.996	0.176	0.151	0.153	0.170
0.17	**0.299**	0.484	0.298	0.162	0.062	0.999	0.483	0.299	0.165	0.073
2.05	**3.056**	0.977	0.922	0.803	0.313	1.000	0.990	0.956	0.877	0.456

Table 2. $E(\pi)$ in percentages, the relative bias of IPFP and ML relative to $E(\pi)$ in percentages based on 10,000 data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.5	22.8	30.8	37.0	36.0	1.8	23.4	32.5	41.3	53.9
0.40	0.6	0.3	0.1	−0.1	0.4	0.7	0.2	−0.0	−0.3	−0.8
0.97	−0.4	0.1	0.6	1.3	3.8	−0.4	0.1	0.6	1.5	5.2
11.8	0.0	0.7	1.2	1.9	4.2	0.0	0.8	1.3	2.1	5.8
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	−2.2	31.6	35.6	35.6	22.6	0.8	34.4	42.1	48.0	54.5
0.40	0.9	−0.2	−0.3	−0.2	0.8	0.7	−0.3	−0.6	−0.8	−1.1
0.97	−0.3	1.1	2.1	3.3	4.4	−0.5	1.1	2.3	3.9	7.7
11.8	0.0	1.6	2.5	3.5	4.6	0.0	1.7	2.8	4.4	9.4
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	1.4	17.3	24.5	31.4	36.1	1.7	17.1	24.7	33.2	49.0
0.40	−0.1	−0.3	−0.5	−0.6	−0.3	−0.3	−0.5	−0.6	−0.7	−1.1
0.97	0.3	0.5	0.8	1.3	3.2	−0.0	0.3	0.6	1.1	4.0
11.8	0.0	0.5	0.9	1.5	3.6	−0.0	0.5	0.9	1.5	4.5
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	−0.7	25.9	28.2	28.9	23.6	1.4	20.7	26.7	33.5	47.4
0.40	1.1	−0.0	−0.1	−0.2	0.1	2.0	0.7	0.3	−0.0	−0.5
0.97	0.4	1.6	2.2	2.8	3.7	0.3	2.1	2.4	2.9	5.0
11.8	−0.0	1.6	2.1	2.6	3.3	0.1	2.1	2.4	3.1	5.7
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	−1.2	15.8	23.4	30.6	36.0	−1.1	15.5	23.3	32.2	49.4
0.40	−1.4	−1.4	−1.4	−1.3	−0.4	−1.7	−1.7	−1.6	−1.46	−1.1
0.97	0.2	0.6	0.9	1.4	3.5	0.2	0.6	0.9	1.5	4.3
11.8	−0.1	0.4	0.8	1.4	3.5	−0.1	0.4	0.8	1.4	4.3
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.1	16.2	23.8	31.7	39.0	1.5	16.3	24.4	33.7	51.4
0.40	−0.4	−0.6	−0.7	−0.8	−0.5	−0.4	−0.6	−0.7	−0.9	−1.2
0.97	−0.4	−0.1	0.2	0.7	3.0	−0.4	−0.1	0.2	0.8	3.7
11.8	−0.0	0.4	0.8	1.3	3.4	−0.0	0.4	0.8	1.4	4.3
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	−5.5	85.1	88.9	81.4	44.2	−5.7	100.5	119.1	131.6	144.1
0.07	−4.3	−32.0	−34.0	−33.7	−23.1	−4.3	−35.8	−41.5	−45.2	−49.0
0.17	5.1	6.0	6.7	7.3	5.6	5.1	6.2	7.4	9.0	12.2
2.05	−0.7	2.2	3.04	3.3	−0.1	−0.7	2.7	4.32	6.0	8.3

of adjusted versions, overall to find a good compromise between bias and efficiency, the methods "+0.5" and "+1" appear good methods and in particular IPFP + 1.

In terms of coverage, the unadjusted version "+0" often suffers from undercoverage. In general, we would expect that adding a constant α to each cell would lead to a decrease in coverage because, due to the artificially increased sample size the standard errors are smaller and the CIs are smaller. The results show, however, that the adjusted versions

Table 3. $E(\pi)$ in percentages, coverage of IPFP and ML methods and their adjusted versions in percentages based on 10,000 simulated data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	41.5	100.0	100.0	100.0	100.0	41.5	100.0	100.0	100.0	100.0
0.40	99.1	99.7	99.9	100.0	100.0	99.1	99.7	99.9	100.0	100.0
0.97	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
11.8	94.9	95.1	95.0	94.5	89.8	95.0	95.2	95.1	94.5	85.9
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	15.8	42.5	42.6	42.5	40.3	14.5	42.6	42.6	42.6	41.5
0.40	80.1	90.8	90.9	91.0	90.7	80.0	90.8	90.9	91.0	90.6
0.97	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7
11.8	97.7	97.7	97.6	96.7	91.9	97.8	97.8	97.5	96.6	80.9
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	29.9	33.2	34.5	34.4	30.5	29.7	33.2	34.8	35.2	32.7
0.40	82.6	84.7	85.5	85.9	84.6	82.7	85.0	85.6	86.1	85.1
0.97	99.3	99.3	99.3	99.4	99.4	99.3	99.3	99.3	99.3	99.3
11.8	89.0	89.6	89.4	88.8	80.1	93.7	93.7	93.5	93.1	84.3
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	19.6	28.7	28.7	27.9	23.8	16.9	28.0	30.5	31.8	31.4
0.40	64.3	80.3	81.2	80.8	78.2	56.9	70.7	74.2	77.7	81.6
0.97	92.9	97.2	97.5	97.4	96.6	86.8	90.0	91.1	92.3	96.1
11.8	63.8	68.5	66.1	62.5	50.9	32.1	34.1	34.5	35.0	35.6
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	29.4	32.9	34.1	34.2	30.2	29.1	32.6	34.4	34.8	32.2
0.40	82.7	84.9	85.6	86.1	84.7	82.0	84.5	85.2	85.8	85.0
0.97	99.2	99.3	99.3	99.3	99.3	98.7	98.9	99.1	99.1	99.3
11.8	89.2	89.0	88.6	87.6	79.7	81.2	81.3	81.3	80.9	74.7
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	30.1	32.2	33.8	34.0	30.2	30.1	32.3	34.0	34.2	31.3
0.40	84.0	85.4	86.0	86.4	85.2	84.0	85.4	86.0	86.4	85.3
0.97	99.2	99.2	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3
11.8	93.1	93.3	93.0	92.7	85.8	94.2	94.1	94.1	93.8	84.7
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	7.6	76.0	76.0	76.0	75.9	7.6	76.0	76.0	76.0	76.0
0.07	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
0.17	65.2	98.4	99.8	100.0	99.9	65.3	98.5	99.9	100.0	99.9
2.05	93.2	92.5	91.8	90.7	89.2	93.1	92.7	91.8	90.1	84.7

"+0.5" and "+1" increase the coverage (improving undercoverage), and only for $\alpha = 2, 10$ the coverage appears to decrease compared to $\alpha = 0.5, 1$. Hence the adjusted versions "+0.5" and "+1" appear to be best, when aiming for the coverage to be near or above 95%. The results are similar to Agresti and Coull (1998) as they show that adding pseudo-observations improves upon coverage, but the results are also dissimilar, because our results rather suggest "+0.5" or "+1" and not "+2".

Figures 1 (three dimensions) and 2 (five dimensions) show boxplots of 10,000 estimates for the smallest and largest cells for ML, ML + 0.5, ML + 1, IPFP, IPFP + 0.5, and IPFP + 1. It shows the effect of reducing the MSE when adjusting the data.

The rejection rate of the GOF tests based on a five percent significance level is represented in Table 4 for the unadjusted data ("+0"), because from the results not presented here, it is clear that adjusting the data ("+α" with $\alpha > 0$) leads to too large type I error under H_0. For IPFP, the adjusted GOF versions G^2_{adj}, W^2_{adj} X^2_{adj} are recommended over the unadjusted versions G^2, W^2 X^2, and for ML and CHI2 all

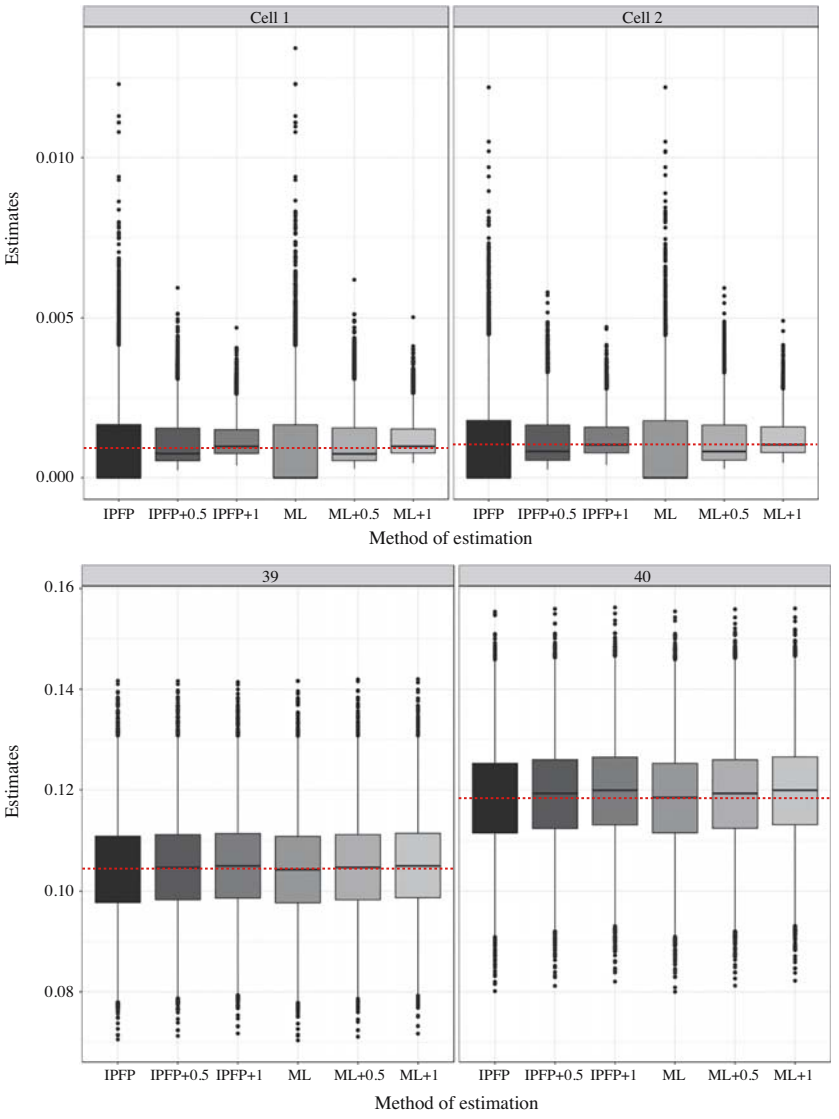

Fig. 1. Boxplots of 10,000 estimates of the methods IPFP, IPFP + 0.5, IPFP + 1, ML, ML + 0.5, and ML + 1 for the two smallest (top) and the two largest (bottom) out of 40 = 5 × 4 × 2 cells in the three dimensional case under random sampling with N = 600 and n = 100 compared with average population proportions (dotted line).

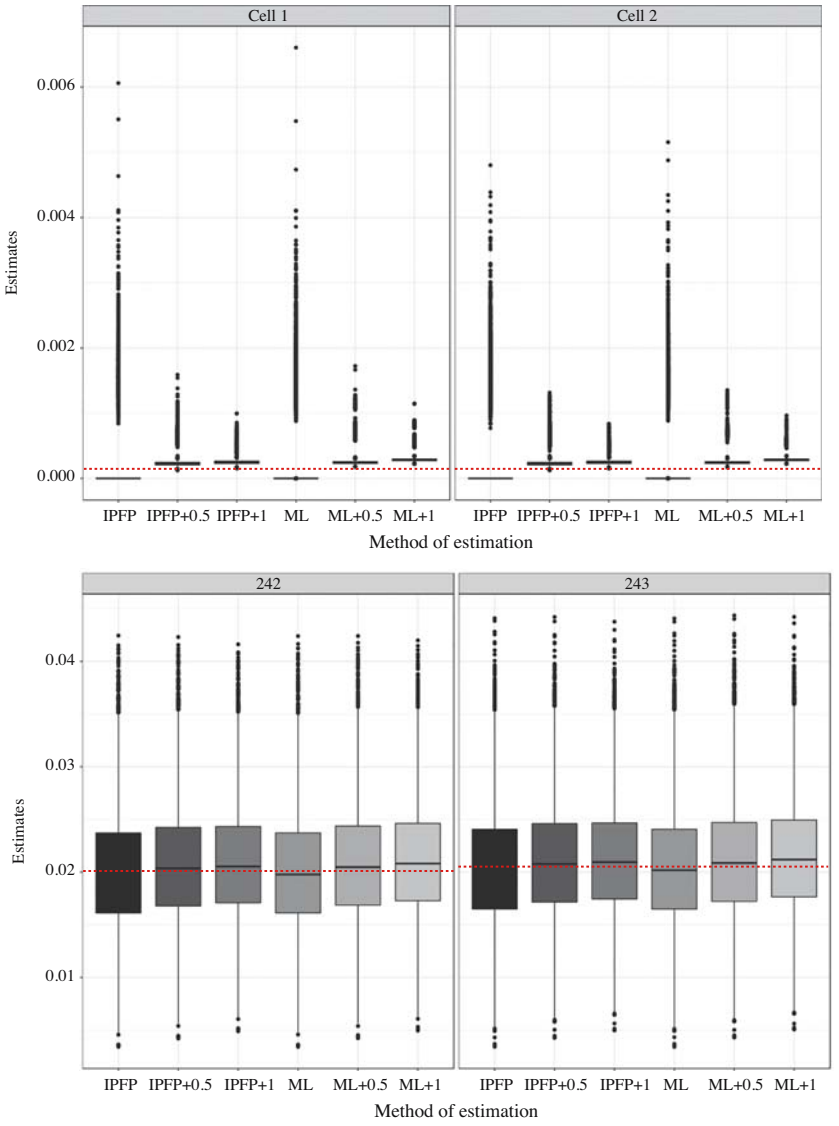

Fig. 2. Boxplots of 10,000 estimates for IPFP, IPFP + 0.5, IPFP + 1, ML, ML + 0.5, and ML + 1 for the two smallest (top) and the two largest (bottom) out of 243 = 3^5 cells in the five dimensional case with three categories in each dimension and under randomsampling with N = 600 and n = 100 compared with average population proportions (dotted line).

GOF versions maintain approximately the type I error, whereas for LSQ the type I error appears too large.

5. Estimating Multidimensional Population Counts for the Illawarra Region

The study area in this article is the Illawarra region in New South Wales, Australia, with a total population of 365,338 in 2011. The Illawarra is the coastal region situated immediately south of Sydney and north of the Shoalhaven or South Coast region

Table 4. Rejection rate of the GOF tests G^2, X^2, W^2 and their adjusted versions (adj) based on $\hat{\pi}$ of the four estimation methods IPFP, ML, CHI2, and LSQ.

	G^2	G^2_{adj}	W^2	W^2_{adj}	X^2	X^2_{adj}
			Dimension = 3, RND, $N = 10{,}000$, $n = 600$			
IPFP	0.143	0.043	0.069	0.068	0.286	0.044
ML	0.041	0.041	0.070	0.068	0.040	0.040
CHI2	0.043	0.043	0.069	0.068	0.038	0.038
LSQ	0.048	0.048	0.070	0.069	0.053	0.053
			Dimension = 3, RND, $N = 600$, $n = 200$			
IPFP	–	0.004	0.019	0.018	–	0.005
ML	0.004	0.004	0.032	0.029	0.003	0.003
CHI2	0.004	0.004	0.019	0.018	0.003	0.003
LSQ	0.095	0.095	0.109	0.091	0.099	0.099
			Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$			
IPFP	1.000	0.993	0.993	0.993	1.000	0.994
ML	0.993	0.993	0.993	0.993	0.993	0.993
CHI2	0.993	0.993	0.992	0.992	0.993	0.993
LSQ	0.994	0.994	0.993	0.993	0.994	0.994
			Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$			
IPFP	–	1.000	1.000	1.000	–	1.000
ML	1.000	1.000	1.000	1.000	1.000	1.000
CHI2	1.000	1.000	1.000	1.000	1.000	1.000
LSQ	1.000	1.000	1.000	1.000	1.000	1.000
			Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$			
IPFP	1.000	0.992	0.992	0.992	1.000	0.992
ML	0.992	0.992	0.992	0.992	0.992	0.992
CHI2	0.992	0.992	0.992	0.992	0.992	0.992
LSQ	0.992	0.992	0.992	0.992	0.993	0.992
			Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$			
IPFP	0.500	0.869	0.873	0.873	0.500	0.871
ML	0.868	0.868	0.873	0.873	0.868	0.868
CHI2	0.869	0.869	0.873	0.872	0.866	0.866
LSQ	0.873	0.873	0.873	0.873	0.876	0.876
			Dimension = 5, RND, $N = 10{,}000$, $n = 600$			
IPFP	–	0.042	0.054	0.050	–	0.039
ML	0.039	0.039	0.054	0.050	0.037	0.037
CHI2	0.042	0.042	0.054	0.050	0.035	0.035
LSQ	0.048	0.048	0.054	0.050	0.050	0.050

(see Figure 3). The smallest geographic area defined in the Australian Statistical Geography Standard (ASGS) is the Statistical Level 1 (SA1), indicated by index j, for which the data are available to our study. The number of males and females living within the study area and three major subregions is presented in Table 5.

The Australian census tables released by the Australian Bureau of Statistics (ABS) available through the ABS Table Builder Pro were used for this study. SA1-specific

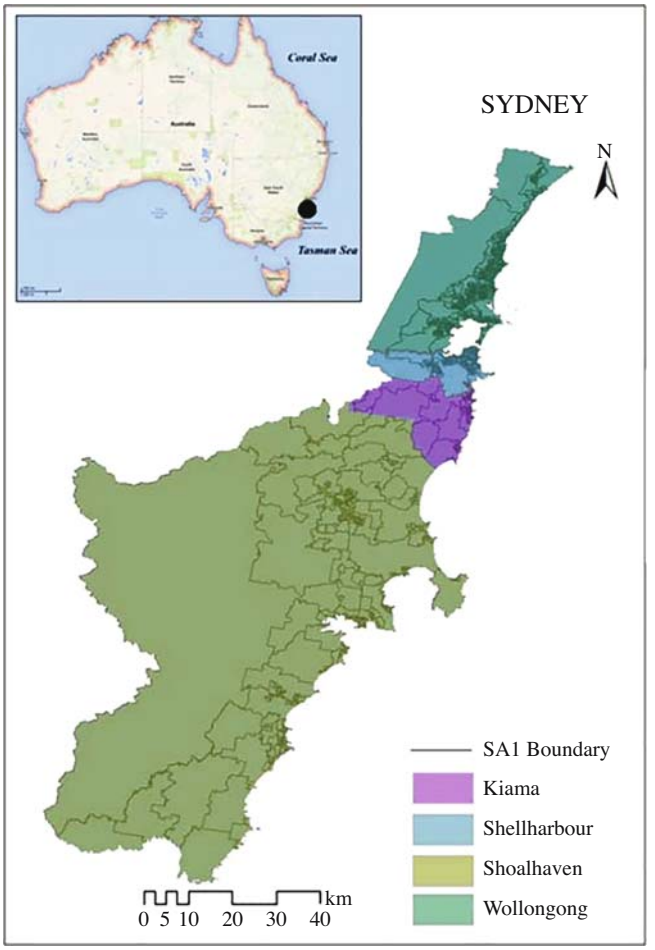

Fig. 3. Map of study area (Illawarra Region).

marginal population counts for age by gender and for family type are available from the census. These tables contain 18 age categories ($0-4, 5-9, \ldots, 80-84, >84$), two genders and four family categories (couple with no children, couple with children, one parent family, other family). Our aim is to find pseudo-census tables for age by sex by family type for each of the SA1 of the Illawarra region.

Table 5. Living population in the study area for the Illawarra and three greater subregions based on 2011 Australian Census Data.

Area	Males	Females	Total
Kiama and Shellharbour	40,160	42,184	82,344
Wollongong	94,986	97,079	192,065
Shoalhaven	44,667	46,262	90,929
Total	179,813	185,252	365,338

There are $144 = 18 \times 2 \times 4$ cells and corresponding probabilities $\pi_{abc}^{(j)}$ for each SA1 $j = 1, \ldots, 61$, and six of these $\pi_{abc}^{(j)}$ are set to zero because 'family couples without children' should have no family member in the age groups $0-4$, $5-9$, $10-14$ for both genders. This leaves 138 cells for each SA1 j.

A 1% Basic Census Sample File (CSF) with $n = 2{,}902$ housing units was available to this study through the Confidentialised Unit Record File (CURF) microdata system. As there is no geographical information (as SA1) attached to the 1% CSF, this sample is used for all of the 61 SA1 study areas with population sizes of $6 \leq N \leq 1060$. As $n > N$, the 1% Basic CSF can only be thought of as a pseudo-sample and not a random sample without replacement from the target-population. The `R` package `mipfp` (Barthélemy and Suesse 2016) is used to produce the raking (IPFP), ML(RS), CHI2, and LSQ estimates.

Figure 4 shows the results when using only the 1% CSF without imposing population constraints. The results do not vary across SA1s, as we have only one sample – the 1% CSF – containing people from the whole Illawarra region, ignoring the available known marginal totals for each SA1 j. Our approach of using this pseudo-sample might seem questionable, as sample and target populations are not the same, but as mentioned in Section 3, IPFP and ML also provide ML estimates under the misspecification models where $\kappa = 0, -1$. Here based on the known marginal totals, these models are of the specific form

$$\left(\frac{\pi_{abc}^{(j)}}{\tau_{abc}}\right)^{\kappa} = \theta^{(j)} + \theta_{1(a)}^{(j)} + \theta_{2(b)}^{(j)} + \theta_{3(c)}^{(j)} + \theta_{12(ab)}^{(j)}; \quad \kappa = -1, 0, 1, 2, \tag{13}$$

where the first variable is age, the second is gender and the third is family type. The specific SA1 is indicated by index j, however it should be noted that $\pi_{abc}^{(j)}$ contains superscript j whereas τ_{abc} does not have superscript j, because the same data set is used as a (pseudo-) sample from a population that is characterized by τ_{abc}. In contrast, each SA1 j has its specific

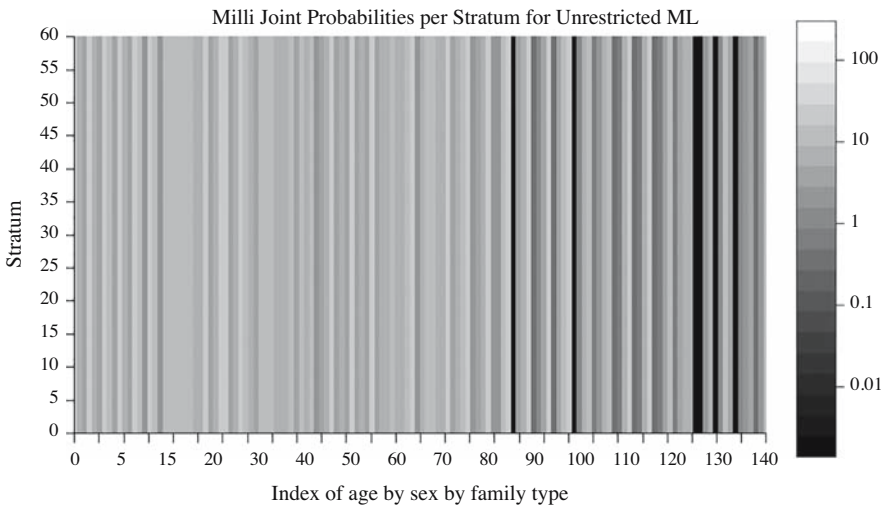

Fig. 4. Unrestricted ML estimator for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file.

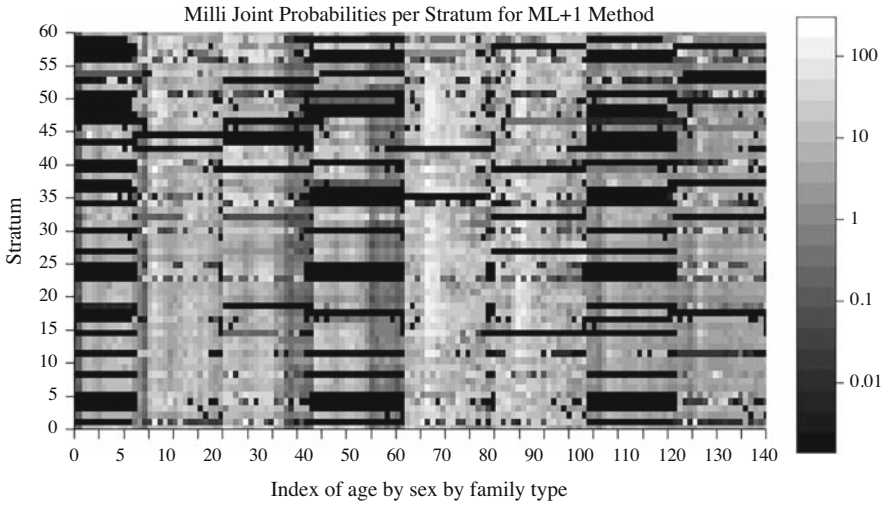

Fig. 5. ML + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

population distribution, denoted by $\pi_{abc}^{(j)}$, and its estimate will be different for each j, due to the availability of known marginal population counts that are specific for SA1 j.

The SA1 could be considered as another variable and the joint distribution containing $61 \times 138 = 8{,}418$ cells could be estimated at once, however this would yield the same results as when estimating 138 cells for each SA1 j separately and would also increase the number of constraints by a factor of 61. Usually the larger the number of cells and the number of constraints become, the more unstable becomes the optimisation algorithm due to the curse of dimensionality. Joint estimation would also complicate the specification of cells and the margins, increasing the chance or errors by the user. Generally, it is not

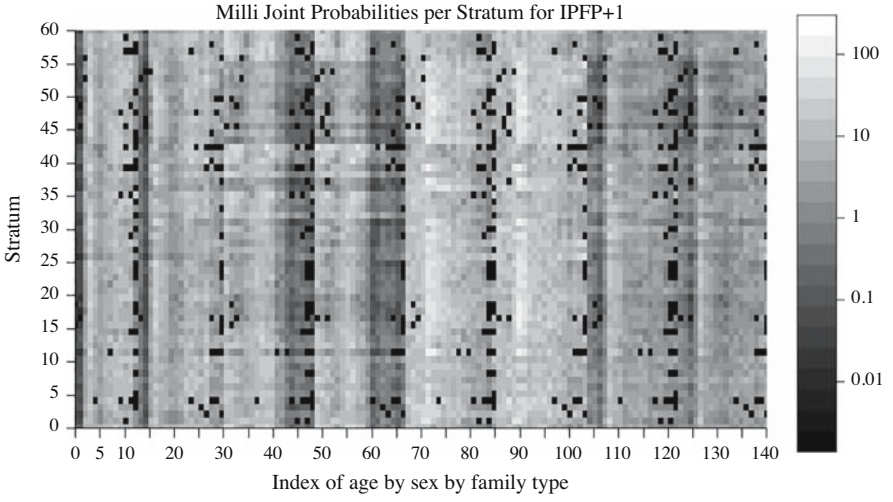

Fig. 6. IPFP + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

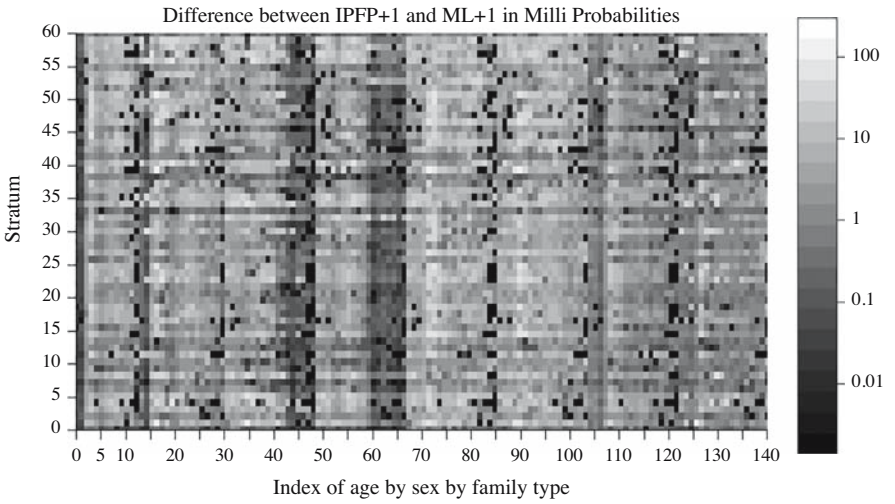

Fig. 7. Absolute differences between ML + 1 and IPFP + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

advisable to increase the dimensionality artificially, if not needed. We rather recommend separate estimation for each SA1 j, using for example the `R` package `mipfp`.

Figure 5 shows the results of the ML + 1 method and Figure 6 shows the results of IPFP + 1 for each geographical area (stratum) referring to 138 probabilities. The results differ, as can be seen in Figure 7. Based on the results of the simulation study and the fact that no true random sample is available, but only a pseudo-sample, IPFP + 1 is preferred to improve efficiency and to improve coverage. If bias is to be avoided, then IPFP + 0 is recommended for the point estimates.

All methods exactly match the population constraints. For example, for SA1 with ID 1114961 the population marginal proportions (relative to SA1 size) for family type are 0.191 (couple with no children), 0.292 (couple with children), 0.410 (one parent family) and 0.107 (other family). Results for IPFP + 0, ML + 0, and the CHI2 and LSQ methods are not shown to preserve space.

The age by sex by family type tables based on a pseudo-sample and two known marginal tables serve as pseudo-census counts/tables, as the true census counts are not released due to confidentiality restrictions. The results are also valuable for population reconstruction, as they form the basis for the simulation of area-specific SPs.

6. Discussion

The main objective of this article is to compare several estimation methods for obtaining population count estimates $\hat{N}_{abc} = N\hat{\pi}_{abc}$ or equivalently, estimates of the joint probabilities $\hat{\pi}_{abc}$, when a sample is available and when marginal population counts (subtables) are known. IPFP, also known as raking, is the standard method to deal with this problem (Ballas et al. 2005; Smith et al. 2005), due primarily to the popularity of IPFP and widely available software. IPFP can also be applied if the seed (sample) is only partially available, for example due to confidentiality restriction.

The ML method is not very popular because of several limitations: The availability of a representative sample is not always warranted, this sample needs to be a true random sample, and there are not many statistical packages that have implemented this approach. On the other hand, IPFP requires non-zero cells of the sample to converge and to provide a unique solution to the underlying optimisation problem. Nonetheless, it must be noted that even with zero cell counts, IPFP often still converges. Another problem with IPFP is that some of the estimated cell counts are zero, when some cells in the marginal population tables are zero, even if a solution exists with positive estimates. This means some combinations of attributes in the simulation process of synthetic populations will be impossible to sample due to the zero estimates. The ML method also has convergence problems when zero cell counts are present.

As an alternative we proposed the data adjustment methods of the form "$+\alpha$" with $\alpha = 0.5, 1, 2, 10$ and the simulation study showed that these methods generally improve efficiency at the costs of increased bias. They also generally improve coverage. Based on all the results, the adjustments "$+0.5$" and "$+1$" appear best. Overall, the simulation study suggests that under the mis-specification models that IPFP $+ 1$ appears to be a reasonable choice in terms of improved coverage, and in terms of efficiency at the costs of increased bias. Under random sampling, instead we recommend ML $+ 1$. Even if biased yet more efficient point estimates might not be desirable for some practitioners, the improved coverage of the confidence intervals is still worth highlighting.

It is sometimes difficult to determine whether random sampling or one of the misspecification models applies and this means that it is difficult to make a decision whether IPFP or ML should be applied. The GOF tests enable testing whether a random sample can be assumed. If the tests are not rejected, then we recommend using ML $+ 1$, and if rejected then IPFP $+ 1$. In some cases, as in our example, it is apparent, for example when $n > N$, that the sample cannot be a true random sample from the target population and the IPFP $+ 1$ method is preferred. The "$+1$" adjustment also solves the possible convergence problem, as now none of the cells is zero. If bias is a major concern, then either the IPFP $+ 0$ or ML $+ 0$ methods should be applied, depending on the results ofthe GOF tests.

IPFP is overall the preferred method under the misspecification models, however estimates of some cells of IPFP might be zero due to zero counts in the marginal tables. If this is undesirable, for example when the synthetic simulation process does not aim at excluding particular combinations of attributes, then the ML method is the preferred alternative (and its adjusted versions), if indeed the ML estimates are non-zero for all cells.

In the SP literature, the presented (co)variance estimators are often unknown and are worth highlighting, as they form the basis of Wald-type confidence intervals, the measure of uncertainty and precision.

Any data set that possesses features that are otherwise not available from the target population is recommended over using artificial data as the seed, as illustrated in Section 5. For example generally, age (by sex) is related to family type, as each family type usually has a particular age (by sex) distribution. While the population tables on age by sex and family type provide information about the marginal distributions, they do not provide information on how age by sex and family type are related. Because no sample was available, Barthélemy and Toint (2013) used equal weights as the seed, however, this will imply incorrectly that there is independence between age by sex and family type, see

Appendix C. In contrast, using some available related samples (even if not a real random sample in the classical sense) that have typical dependence between age by sex and family type present will preserve the relationship between age by sex and family type in the final estimates. This preservation of higher order terms in the respective log-linear model (Mosteller 1968), when IPFP is applied, is clearly advantageous. This is also advantageous in the classical sense. The seed is one data set and the marginal population tables form another data set. Using the two real data sets jointly will provide more information than a single real data set alone.

In this article, we only considered four misspecification models. In practice, however, this class of models might be too narrow to obtain accurate estimates in all cases. Developing a wider class of misspecification models and the investigation of the best performing method under this extended class will be the subject of future research.

Appendix A. Form of Maximum Likelihood Estimates

Let us write the constrained log-likelihood L_c, see (5), with second order population constraints as

$$L_c = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc} + \sum_a \sum_b \lambda_{a,b}(\pi_{ab\bullet} - (N_{ab\bullet}/N))$$

$$+ \sum_a \sum_c \lambda_{a,c}(\pi_{a\bullet c} - (N_{a\bullet c}/N)) + \sum_b \sum_c \lambda_{b,c}(\pi_{\bullet bc} - (N_{\bullet bc}/N)).$$

Now let us take first derivatives with respect to π_{abc}

$$\frac{\partial L_c}{\partial \pi_{abc}} = \frac{y_{abc}}{\pi_{abc}} - \lambda_{a,b} - \lambda_{a,c} - \lambda_{b,c},$$

where $\lambda_{a,b}$, $\lambda_{a,c}$ and $\lambda_{b,c}$ are Lagrange multiplier determined by the ML algorithm.

Setting derivatives to zero $\frac{\partial L_c}{\partial \pi_{abc}} = 0$ and imposing typical constraints such as second order parameters sum to zero, estimates have the form

$$\left(\frac{\hat{\pi}^{ML}_{abc}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}^{ML}_{1(a)} + \hat{\theta}^{ML}_{2(b)} + \hat{\theta}^{ML}_{3(c)} + \hat{\theta}^{ML}_{12(ab)} + \hat{\theta}^{ML}_{13(ac)} + \hat{\theta}^{ML}_{23(bc)}. \tag{A.1}$$

It also shows that if second order population constraints are included, then the form of the estimates include second order terms. If for example first order population constraints are included, then the right hand side of (A.1) will only contain main effects (first order terms).

Appendix B. Showing that IPFP Estimates are ML Estimates under Model (12) with $\kappa \to 0$

Suppose sampling fractions are small, then y_{abc} are approximately multinomially distributed and the sample proportions p_{abc} are ML estimates of τ_{abc}. By Model (12) with $\kappa \to 0$, the population probabilities π_{abc} are given by

$$\pi_{abc} = \tau_{abc} \exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)}),$$

and ML estimates of the θ's are obtained by solving

$$\pi_{ab\bullet} = \sum_c \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{a\bullet c} = \sum_b \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{\bullet bc} = \sum_a \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)}).$$

As the ML estimates of functions of τ_{abc} are the functions evaluated at $\hat{\tau}_{abc} = p_{abc}$, the ML estimates of π_{abc} are of the form

$$\hat{\pi}_{abc} = p_{abc}\exp(\hat{\theta} + \hat{\theta}_{1(a)} + \hat{\theta}_{2(b)} + \hat{\theta}_{3(c)} + \hat{\theta}_{12(ab)} + \hat{\theta}_{13(ac)} + \hat{\theta}_{23(bc)}),$$

where the $\hat{\theta}$ estimates are obtained by solving

$$\pi_{ab\bullet} = \sum_c p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{a\bullet c} = \sum_b p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{\bullet bc} = \sum_a p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

These equations are solved by the raking estimates, see Equation (4).

Similar arguments can be shown to show that MLRS provides ML estimates for Model (12) with $\kappa = -1$, LSQ provides ML estimates for Model (12) with $\kappa = 1$ and CHI2 provides ML estimates for Model (12) with $\kappa = 2$.

Appendix C. Independence with Equal Weights

Suppose we have three variables and suppose equal initial weights as the seed, i.e., $\pi^{(0)}_{abc} \propto 1$, which implies that $\pi^{(0)}_{abc} = \frac{1}{\sum_{abc} 1} = \frac{1}{K}$ $(K = ABC)$.

$$\pi^{(0)}_{ab\bullet} = \sum_c \pi^{(0)}_{abc} = \frac{C}{K} = \frac{1}{AB}$$

$$\pi^{(0)}_{a\bullet c} = \sum_b \pi^{(0)}_{abc} = \frac{B}{K} = \frac{1}{AC}$$

$$\pi^{(0)}_{\bullet bc} = \sum_a \pi^{(0)}_{abc} = \frac{A}{K} = \frac{1}{BC}$$

$$\pi^{(0)}_{a\bullet\bullet} = \sum_{b,c} \pi^{(0)}_{abc} = \frac{BC}{K} = \frac{1}{A}$$

$$\pi^{(0)}_{\bullet b\bullet} = \sum_{a,c} \pi^{(0)}_{abc} = \frac{AC}{K} = \frac{1}{B}$$

$$\pi^{(0)}_{\bullet\bullet c} = \sum_{a,b} \pi^{(0)}_{abc} = \frac{AB}{K} = \frac{1}{C}$$

Now from these equations, it is apparent that the three categorical variables are independent when $\pi^{(0)}_{abc}$ would be the final estimates (zero iterations of IPFP).

Assuming that population counts are available for each of the three variables, it should be noted that the raking/IPFP estimates denoted by $\hat{\pi}^r_{abc}$ are of the following form (Little and Wu 1991)

$$\log\left(\frac{\hat{\pi}^r_{abc}}{p_{abc}}\right) = \hat{\theta}^r + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)},$$

similar to Equation (4), where $p_{abc} = \frac{1}{K}\left(= \pi^{(0)}_{abc}\right)$, because the "sample" consists of equal weights (pseudo-data), as no real data set/sample is available. Hence

$$\log\left(\hat{\pi}^r_{abc}\right) = \text{const} + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)}$$

and it follows that

$$\hat{\pi}^r_{abc} = \frac{1}{g_{\bullet\bullet\bullet}}\exp\left(\hat{\theta}^r_{1(a)}\right) \times \exp\left(\hat{\theta}^r_{2(b)}\right) \times \exp\left(\hat{\theta}^r_{3(c)}\right) = \frac{1}{g_{\bullet\bullet\bullet}}\alpha_a \times \alpha_b \times \alpha_c,$$

where $g_{\bullet\bullet\bullet} = \sum_{a,b,c}\alpha_a\alpha_b\alpha_c = \left[\sum_a \alpha_a\right] \times \left[\sum_b \alpha_b\right] \times \left[\sum_c \alpha_c\right]$. From this we obtain the estimated marginal probabilities

$$\hat{\pi}^r_{a\bullet\bullet} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{b,c}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{b,c}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_a}{\sum_a \alpha_a}$$

$$\hat{\pi}^r_{\bullet b\bullet} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{a,c}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{a,c}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_b}{\sum_b \alpha_b}$$

$$\hat{\pi}^r_{\bullet\bullet c} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{a,b}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{a,b}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_c}{\sum_c \alpha_c}$$

and therefore we conclude independence, because, for example, the following equation holds

$$\hat{\pi}^r_{abc} = \hat{\pi}^r_{a\bullet\bullet} \times \hat{\pi}^r_{\bullet b\bullet} \times \hat{\pi}^r_{\bullet\bullet c}.$$

When, for example, X_1 and X_2 are age and sex and X_3 is family type and the known population margins are provided for age by sex (i.e., (X_1, X_2) known) and for family type (X_3 known), then similarly final estimates will imply that still (X_1, X_2) and X_3 are independent.

Appendix D. Simulation Results of CHI2 and LSQ

Table D.6. $E(\pi)$ in percentages, RMSE for methods CHI2 and LSQ relative to IPFP + 0, RMSE $<$ 1 indicates better and RMSE $>$ 1 worse, all based on 10,000 simulated data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.038	0.362	0.244	0.193	0.213	0.997	0.368	0.244	0.178	0.388
0.40	1.010	0.662	0.479	0.293	0.076	1.000	0.673	0.497	0.325	0.175
0.97	0.993	0.850	0.738	0.577	0.203	1.006	0.860	0.748	0.587	0.219
11.8	0.995	0.969	0.959	0.964	1.380	1.117	0.980	0.970	0.959	0.785
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	1.099	0.294	0.316	0.361	0.439	0.941	0.314	0.347	0.417	0.620
0.40	1.021	0.375	0.288	0.274	0.354	1.034	0.397	0.320	0.318	0.427
0.97	0.999	0.645	0.493	0.374	0.379	1.390	0.664	0.513	0.396	0.376
11.8	0.987	0.923	0.911	0.941	1.673	6.633	0.948	0.917	0.858	0.593
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	1.105	0.653	0.688	0.861	1.514	1.073	0.866	1.039	1.441	2.360
0.40	1.068	0.757	0.651	0.610	0.962	1.182	0.975	0.915	0.983	2.071
0.97	1.049	0.873	0.775	0.661	0.652	1.363	1.200	1.101	0.991	1.213
11.8	0.882	0.873	0.875	0.894	1.333	2.985	1.809	1.784	1.792	2.196
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	1.441	0.572	0.485	0.454	0.535	1.287	0.998	1.229	1.458	1.442
0.40	1.786	0.912	0.702	0.524	0.335	1.472	0.798	0.819	0.974	1.559
0.97	2.918	1.926	1.613	1.247	0.515	1.753	1.022	0.867	0.808	1.060
11.8	5.900	4.884	4.823	4.705	4.003	6.221	3.474	3.104	2.715	2.477
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	1.222	0.754	0.734	0.901	1.606	1.007	0.795	1.031	1.579	2.898
0.40	1.471	1.032	0.832	0.693	0.952	1.026	0.806	0.761	0.862	2.122
0.97	1.928	1.628	1.428	1.145	0.750	1.067	0.886	0.817	0.757	1.126
11.8	2.553	2.501	2.472	2.431	2.466	1.893	0.850	0.869	0.956	1.782
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.072	0.707	0.767	1.006	1.911	1.007	0.752	0.857	1.185	2.315
0.40	0.999	0.788	0.700	0.662	1.090	1.036	0.845	0.770	0.763	1.431
0.97	0.988	0.895	0.826	0.736	0.733	1.091	0.958	0.890	0.802	0.864
11.8	0.916	0.901	0.899	0.916	1.401	1.855	1.131	1.128	1.146	1.357
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	0.997	0.221	0.224	0.246	0.282	1.007	0.116	0.091	0.147	0.151
0.07	0.994	0.178	0.156	0.160	0.178	0.992	0.183	0.118	0.110	0.683
0.17	0.998	0.484	0.302	0.168	0.074	0.999	0.484	0.297	0.162	0.064
2.05	1.000	1.000	0.983	0.933	0.595	1.001	0.960	0.880	0.732	0.488

Table D.7. $E(\pi)$ in percentages, the relative bias of CHI2 and LSQ relative to $E(\pi)$ in percentages based on 10,000 data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.7	23.5	33.2	42.7	56.2	2.4	22.0	27.4	24.8	−69.9
0.40	0.7	0.2	−0.1	−0.4	−1.2	0.6	0.3	0.2	0.3	3.9
0.97	−0.4	0.1	0.6	1.6	5.9	−0.4	0.1	0.5	1.2	2.3
11.8	0.0	0.7	1.3	2.3	7.0	0.1	0.7	1.2	1.7	1.5
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	0.2	35.5	43.9	50.2	57.1	2.0	24.7	12.9	−20.9	−98.9
0.40	0.7	−0.3	−0.7	−1.1	−1.8	−0.2	0.1	0.4	1.6	6.2
0.97	−0.4	1.1	2.5	4.4	8.7	0.3	1.0	1.8	2.6	0.8
11.8	0.0	1.8	3.1	5.0	12.5	−2.0	1.5	2.0	2.1	−2.7
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	2.0	17.0	24.8	34.0	51.3	2.4	17.6	24.5	28.5	−9.4
0.40	−0.1	−0.1	−0.6	−0.8	−1.4	0.0	−0.0	−0.2	−0.2	1.5
0.97	−0.1	0.2	0.4	1.1	4.5	0.6	0.8	1.1	1.5	2.4
11.8	0.1	0.5	0.8	1.5	4.9	−0.0	0.5	0.9	1.4	2.2
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	2.4	19.4	26.1	34.2	50.5	4.9	32.9	34.4	28.8	−16.1
0.40	2.0	0.71	0.2	−0.1	−0.2	0.6	−1.4	−1.3	−0.6	3.9
0.97	0.4	2.2	2.5	2.9	5.2	−0.3	1.3	1.9	2.5	2.1
11.8	0.3	2.4	2.7	3.2	6.2	1.9	1.8	2.2	2.4	0.6
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	−1.1	15.4	23.2	32.9	51.8	−0.7	15.8	23.0	27.5	−5.4
0.40	−1.9	−1.8	−1.7	−1.6	−1.4	−1.3	−1.3	−1.3	−1.1	1.5
0.97	0.4	0.6	0.9	1.4	4.7	0.4	0.7	1.0	1.4	2.6
11.8	0.1	0.5	0.8	1.4	4.7	−0.1	0.5	0.9	1.4	2.1
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.7	16.3	24.7	34.5	53.7	1.5	16.1	23.0	27.4	−13.8
0.40	−0.3	−0.6	−0.8	−0.9	−1.4	−0.5	−0.6	−0.7	−0.8	1.0
0.97	−0.3	−0.1	0.2	0.8	4.2	−0.4	−0.1	0.2	0.6	2.2
11.8	−0.0	0.4	0.8	1.4	4.9	−0.083	0.4	0.7	1.2	2.0
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	−5.9	106.2	126.8	140.4	153.5	−4.8	43.0	−25.8	−94.8	−97.2
0.07	−4.4	−37.2	−43.1	−46.9	−50.5	−4.6	−22.2	−9.3	20.1	115.6
0.17	5.1	6.5	7.9	9.7	12.4	5.1	5.7	5.8	5.7	2.7
2.05	−0.71	3.1	5.2	7.7	12.3	−0.68	1.4	1.0	−1.2	−13.2

Table D.8. $E(\pi)$ in percentages, coverage of CHI2 and LSQ methods and their adjusted versions in percentages based on 10,000 simulated data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	41.3	99.9	99.9	99.9	99.9	41.5	99.9	99.9	99.8	42.6
0.40	99.0	99.6	99.9	100.0	100.0	99.0	99.6	99.8	100.0	100.0
0.97	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
11.8	95.0	95.2	95.2	94.5	81.2	94.7	94.9	94.8	94.2	92.0
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	14.3	42.5	42.6	42.6	41.5	17.2	41.6	39.1	30.2	0.9
0.40	79.9	90.7	90.9	90.9	90.6	79.4	90.7	90.9	90.9	90.4
0.97	99.6	99.6	99.6	99.6	99.6	98.1	99.6	99.6	99.6	99.6
11.8	97.8	97.8	97.5	96.3	68.3	90.4	97.6	97.4	96.6	91.7
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	29.5	33.0	34.7	35.1	32.7	29.7	31.8	31.4	28.8	16.0
0.40	82.4	84.7	85.6	85.9	84.9	81.4	83.0	83.8	83.7	77.6
0.97	99.0	99.2	99.3	99.3	99.3	98.8	99.1	99.1	99.2	98.7
11.8	90.6	90.7	90.5	89.8	78.2	79.1	79.8	79.5	78.5	66.7
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	16.1	26.5	29.4	31.4	31.3	15.3	20.3	17.5	13.6	6.9
0.40	53.5	66.6	70.7	75.1	80.8	51.8	63.7	64.1	61.0	44.8
0.97	83.3	87.2	88.3	90.2	95.0	80.5	89.1	91.2	92.1	85.9
11.8	25.7	27.9	28.4	28.7	29.5	28.9	31.3	31.6	31.3	23.3
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	28.6	32.1	34.1	34.7	32.3	29.5	32.1	31.6	28.6	14.8
0.40	81.0	83.7	84.7	85.5	84.8	82.4	84.6	85.2	85.0	77.7
0.97	97.9	98.2	98.4	98.8	99.2	98.9	99.2	99.2	99.2	99.0
11.8	73.5	73.9	74.0	74.0	68.7	90.2	90.9	90.4	88.1	70.8
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	29.9	32.1	33.9	34.2	31.3	30.1	32.0	33.2	31.9	17.2
0.40	84.0	85.4	85.9	86.4	85.2	83.7	85.1	85.5	86.0	83.3
0.97	99.2	99.2	99.2	99.2	99.2	99.1	99.2	99.2	99.2	99.2
11.8	94.1	94.2	94.1	93.6	81.7	90.8	91.4	91.0	90.4	82.5
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	7.7	76.0	76.0	76.0	76.0	7.7	75.9	65.3	4.8	0.0
0.07	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
0.17	65.3	98.5	100.0	100.0	99.9	65.2	98.3	99.8	100.0	99.8
2.05	93.1	92.9	91.8	89.8	79.2	93.3	92.2	91.6	89.8	72.0

7. References

Agresti, A. 2002. *Categorical Data Analysis*. New York: Wiley.

Agresti, A. and B.A. Coull. 1998. "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions." *The American Statistician* 52: 119–126. Doi: http://dx.doi.org/10.1080/00031305.1998.10480550.

Agresti, A. and D.B. Hitchcock. 2005. "Bayesian Inference for Categorical Data Analysis." *Statistical Methods and Applications* 14: 297–330. Doi: http://dx.doi.org/10.1007/s10260-005-0121-y.

Arentze, T., H. Timmermans, and F. Hofman. 2007. "Creating Synthetic Household Populations: Problems and Approach." *Journal of the Transportation Research Board, 2014*, 85–91. Doi: http://dx.doi.org/10.3141/2014-11.

Ballas, D., G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter. 2005. "Simbritain: A Spatial Microsimulation Approach to Population Dynamics." *Population, Space and Place* 11: 13–34. Doi: http://dx.doi.org/10.1002/psp. 351.

Barthélemy, J. and T. Suesse. 2016. "mipfp: Multidimensional Iterative Proportional Fitting and Alternative Models. R package version 3.1." Available from: http://CRAN.R-project.org/package=mipfp.

Barthélemy, J. and P.L. Toint. 2013. "Synthetic Population Generation without a Sample." *Transportation Science* 47: 266–279. Doi: http://dx.doi.org/10.1287/trsc.1120.0408.

Beckman, R., K. Baggerly, and M. McKay. 1996. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice* 30: 415–429. Doi: http://dx.doi.org/10.1016/0965-8564(96)00004-3.

Bergsma, W., M. Croon, and J. Hagenaars. 2009. *Marginal Models for Dependent, Clustered and Longitudinal Categorical Data*. New York: Springer.

Causey, B.D. 1983. "Estimation of Proportions for Multinomial Contingency Tables Subject to Marginal Constraints." *Communications in Statistics-Theory and Methods* 12: 2581–2587. Doi: http://dx.doi.org/10.1080/03610928308828624.

De Campos, C.P. and A. Benavoli. 2011. "Inference with Multinomial Data: Why to Weaken the Prior Strength." In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain July 16–22, 2011, Volume 22, pp. 2107. Available at: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/3292.

Deming, W. and F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known." *Annals of Mathematical Statistics* 11: 367–484. Available at: http://www.jstor.org/stable/2235722.

Deville, J., C. Särndal, and O. Sautory. 1991. "Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 86: 87–95.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flotterod. 2013. "Simulation Based Population Synthesis." *Transportation Research Part B: Methodological* 58: 243–263. Doi: http://dx.doi.org/10.1016/j.trb.2013.09.012.

Fienberg, S. 1970. "An Iterative Procedure for Estimation in Contingency Tables." *Annals of Mathematical Statistics* 41: 907–917. Available at: http://www.jstor.org/stable/2239244.

Gange, S.J. 1995. "Generating Multivariate Categorical Variates Using the Iterative Proportional Fitting Algorithm." *American Statistician* 49: 134–138. Available at: http://www.tandfonline.com/doi/abs/10.1080/00031305.1995.10476130.

Gargiulo, F., S. Ternes, S. Huet, and G. Deffuant. 2010. "An Iterative Approach for Generating Statistically Realistic Populations of Households." *PLOS ONE* 5(1), e8828. Doi: http://dx.doi.org/10.1371/journal.pone.0008828.

Geard, N., J. McCaw, A. Dorin, K. Korb, and J. McVernon. 2013. "Synthetic Population Dynamics: A Model of Household Demography." *Journal of Artificial Societies and Social Simulation* 16(1): 1–23. Doi: http://dx.doi.org/10.18564/jasss.2098.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2003. *Bayesian Data Analysis* (2nd ed.). New York: Chapman and Hall/CRC Press.

Harland, K., A. Heppenstall, D. Smith, and M. Birkin. 2012. "Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques." *Journal of Artificial Societies and Social Simulation* 15: 1–24. Doi: http://dx.doi.org/10.18564/jasss.1909.

Huynh, N., J. Barthelemy, and P. Perez. 2016. "A Heuristic Combinatorial Optimisation Approach to Synthesising a Population for Agent Based Modelling Purposes." *Journal of Artificial Societies and Social Simulation* 19: 11. Doi: http://dx.doi.org/10.18564/jasss.3198.

Ireland, C. and S. Kullback. 1968. "Contingency Tables with Given Marginals." *Biometrika* 55: 179–199. Doi: https://doi.org/10.1093/biomet/55.1.179.

Jeffeys, H. 1998. *The Theory of Probability*. Oxford: Oxford University Press.

Lang, J. 1996. "Maximum Likelihood Methods for a Generalized Class of Loglinear Models." *Annals of Statistics* 24: 726–752.

Lang, J. 2004. "Multinomial-Poisson Homogeneous Models for Contingency Tables." *Annals of Statistics* 32: 340–383.

Lang, J. 2005. "Homogeneous Linear Predictor Models for Contingency Tables." *Journal of the American Statistical Association* 100: 121–134. Doi: http://dx.doi.org/10.1198/016214504000001042.

Lang, J. and A. Agresti. 1994. "Simultaneously Modelling Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the American Statistical Association* 89: 625–632.

Lenormand, M. and G. Deffuant. 2013. "Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods." *Journal of Artificial Societies and Social Simulation* 16: 1–16. Doi: http://dx.doi.org/10.18564/jasss.2319.

Little, J. and M. Wu. 1991. "Models for Contingency Tables with Known Margins when Target and Sampled Population Differ." *Journal of the American Statistical Association* 86: 87–95.

Lu, H. and A. Gelman. 2003. "A Method for Estimating Design-Based Sampling Variances for Surveys with Weighting, Poststratification, and Raking." *Journal of Official Statistics* 19: 133–151.

Mosteller, F. 1968. "Association and Estimation in Contingency Tables." *Journal of the American Statistical Association* 63: 1–28. Doi: http://dx.doi.org/10.2307/2283825.

Purcell, N. and L. Kish. 1980. "Postcensal Estimates for Local Areas (or Domains)." *International Statistical Review* 43: 3–18.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Smith, D., G. Clarke, and K. Harland. 2005. "Improving the Synthetic Data Generation Process in Spatial Microsimulation Models." *Environment and Planning* A 41: 1251–1268. Doi: https://doi.org/10.1068/a4147.

Smith, J. 1947. "Estimation of Linear Functions of Cell Proportions." *Annals of Mathematical Statistics* 18: 231–254.

Stephan, F. 1942. "Iterative Method of Adjusting Frequency Tables when Expected Margins are Known." *Annals of Mathematical Statistics* 13(2): 166–178.

Zhang, L. and R. Chambers. 2004. "Small Area Estimates for Cross-Classifications." *Journal of the Royal Statistical Society:* Series B 66: 479–496. Doi: http://dx.doi.org/10.1111/j.1369-7412.2004.05266.x.

Received March 2015
Revised February 2017
Accepted February 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1051–1085, http://dx.doi.org/10.1515/JOS-2017-0049

Figuring Figures: Exploring Europeans' Knowledge of Official Economic Statistics

Maria R. Vicente[1] *and Ana J. López*[1]

Economic issues have been a major concern for Europeans in the last few years. In this context, it is reasonable to suppose that people are aware of the main economic figures regarding Europe. But are they really familiar with them? Do they know what the rates of growth, unemployment and inflation are?

The aim of this article is to explore the level of knowledge of these three economic indicators among Europeans. Several regression models are specified and estimated in order to identify the relationship between an individual's knowledge and their socioeconomic profile, use of the Internet, perceived importance of economic issues and official statistics and trust in them. Cross-country differences are also assessed.

Key words: Economic indicators; European Union; literacy; misperception.

1. Introduction

In June 2010, the United Nations General Assembly adopted 20 October 2010 as the first World Statistics Day (United Nations General Assembly 2010). Such a declaration aimed to acknowledge the importance of official statistics as an indispensable element for both individual and collective informed decision-making (OECD 2005; United Nations General Assembly 2010). Reliable and objective statistics are the basis for democratic societies to function properly: not only are statistics a key input for policy-makers at all levels (European, national, local) but they also serve the public by providing an accurate picture of the current economy and society (Rose 1991; Gal 2002; Ottaviani 2002; Wild 2005; Holt 2008; Eurostat 2016a). Despite their fundamental importance, there is a major concern about the level of public knowledge of official statistics. Research in the United States (US) has shown that the level of knowledge is not high (Blendon et al. 1997; Blinder and Krueger 2004; Curtin 2008, 2009): even though 58% of US adults report knowing the latest rate of unemployment, less than one quarter (25%) indicate that they are familiar with the latest figures of Gross Domestic Product (GDP) growth or inflation (Curtin 2009). In Europe, the empirical evidence is limited (Papacostas 2008; Giovannini et al. 2015 for Italy) and many times the term 'knowledge' refers to being aware of the national statistical

[1] University of Oviedo – Applied Economics, Campus del Cristo s/n, Oviedo 33006, Spain. Emails: mrosalia@uniovi.es and anaj@uniovi.es

Acknowledgments: The authors would like to thank the editor and referees for their comments on previous versions of the manuscript. We also thank for the comments made by participants at the session on Statistical Literacy at the European Conference on Quality in Official Statistics, held in Madrid in June 2016.

offices rather than actual knowledge of official figures (Natcen 2015; Northern Ireland Statistics and Research Agency 2015).

In this context, this article tries to contribute to bridging this gap in research by providing some evidence of Europeans' knowledge of official economic statistics. Results can provide useful insights to identifying those groups which know little about economic statistics, and thus to design appropriate measures to improve general knowledge in society. In particular, attention is paid to the figures of GDP, unemployment and inflation.

2. Background

The role of information in markets and agents' decisions has been a major issue in the economic literature. Neoclassical economists consider that individuals are not only rational but they also have perfect information about the relevant conditions of the economy.

Later theories and models have criticized severely these assumptions and led to the introduction of incomplete and asymmetrical information into economic models. It is then argued that in market transactions, agents have limited information (Lucas 1972; Townsend 1983); additionally, one of the parties might know more than the other (Akerlof 1970).

In this context, rational inattention and sticky information models state that information acquisition involves some costs (Sims 2003), including the time and money of obtaining, processing, and analyzing information and deciding how to use it (Reis 2006). Moreover, costs might vary among individuals, being substantial for those who either do not know how and where to obtain information or do not have the skills to process and understand it (Sims 2003; Blinder and Krueger 2004). Therefore, it is postulated that, for some individuals, the costs might exceed the perceived benefits derived from information; consequently, they would rationally choose not to update (Mankiw and Reis 2002; Sims 2003; Bacchetta and van Wincoop 2005; Reis 2006). Additionally, some authors have hypothesized that individuals' demand for information might differ depending on the sources checked, the volume of news and their type (Akerlof et al. 2000; Souleles 2001; Carroll 2003; Blinder and Krueger 2004; Curtin 2009). In this sense, Kahneman and Tversky (1979) suggest that agents are more receptive towards bad news than good news.

Modern theories have also introduced information heterogeneity into economic models. Thus, individuals might base their decisions on different sets of information. In particular, it is considered that the relevant information for individuals' decisions varies both across people and time depending on their personal circumstances and interests (Blendon et al. 1997; Bryan and Venkatu 2001; Souleles 2001; Blinder and Krueger 2004; Curtin 2008, 2009). An individual might find unemployment figures in his region and sector of activity more meaningful than the national unemployment rate (Curtin 2008; Cardoso et al. 2016).

In this context, empirical evidence has shown that much of the population appears to be rather inattentive to the information stemming from official statistics. Moreover, there are significant gaps of knowledge across socioeconomic groups (Blendon et al. 1997; Walstad 1997; Walstad and Rebeck 2002; Blinder and Krueger 2004; Curtin 2008, 2009; Papacostas 2008; Giovannini et al. 2015).

3. Data Description

The data used in this article come from the Eurobarometer 83.3 carried out on behalf of the European Commission (2015a) in May 2015, and cover the population of individuals aged 15 years and older living in one of the 28 Member States of the European Union. A multi-stage, stratified random sample design was applied in each country in order to guarantee that the sample drawn was representative of the population. Fieldwork ran from May 16 to May 27, 2015. A total 27,758 interviews were successfully made. Among these, 27,745 individuals provided information on their knowledge of official statistics, specifically, of the rates of GDP growth, inflation and unemployment. Eurostat (2016b) defines Gross Domestic Product as a measure of the economic activity that takes into account the value of all goods and services produced, minus the value of the goods or services used in their production. The inflation rate is defined as the percentage change in the price level in a certain period (Eurostat 2016c). The unemployment rate refers to the proportion of unemployed people as part of the total labor force; an unemployed person is defined as a person aged 15–74 years old who does not have a job during the reference week, is actively seeking work, and is available to start working within the next two weeks (Eurostat 2016d).

Sampled individuals were asked about such indicators in the following terms:

"What was the official growth rate of the economy (measured in terms of Gross Domestic Product) in your country in 2014? I can tell you that this figure is between −5% and 15%.

What was the official inflation rate, the rate at which consumer prices increased or decreased, in your country in 2014? I can tell you that the exact figure is between −5% and 20%.

Do you think that, in your country, the inflation rate in 2014 was higher, lower or equal to the rate in 2013?

What was the official unemployment rate, the percentage of active people who do not have a job, in your country in 2014? I can tell you that the exact figure is between 0% and 30%" (European Commission 2015a, 35–36).

Answering these questions might place different cognitive burdens on respondents. As previous research notes (Curtin 2008, 2009), the unemployment rate is the most popular and the easiest indicator to remember, since it is a simple proportion. In contrast, the inflation rate is quite a complex figure to remember: while the rise (decrease) of prices has a direct impact on people's life, recalling its rate of change is difficult since several figures are usually reported at the same time (monthly and annual rates, general and core inflation). In the same sense, the growth rate is a difficult figure to remember: on the one hand, the public is less familiar with the measure of GDP; on the other hand, it is reported quarterly and annually, in nominal and real terms, and its figures are usually revised.

Moreover, it is important to note that the questions provide some guidelines for the answers. In particular, the questionnaire indicates that the rates go from −5% to +15% for growth, from −5% to +20% for inflation and from 0% to +30% for unemployment. The provision of this kind of guidelines typically has some effect on responses. On the one hand, it might prevent people from reporting numbers that make little sense. This would contribute, at least to some extent, to the accuracy of responses. In the particular case of growth and inflation, these guidelines might help respondents to be aware of potential

negative rates that are less intuitive than positive figures. On the other hand, these guidelines could alter people's responses, diverting them from the actual figures they recall. Whether the provision of these guidelines has a net positive or negative effect on accuracy is difficult to assess: they might help some respondents to provide more accurate figures, while for others it could have the opposite effect. Unfortunately, the survey does not have enough information to be able to disentangle such effects. Additionally, we should consider that the provided guidelines are the same for all member countries. Since there is substantial variation in economic figures across nations, some evidence of these cross-country differences may be found in the country constants when running regressions.

Table 1 shows the percentages of 'don't know' answers for the questions on the economic indicators. As just indicated, growth and inflation rates seem to be the most difficult figures to remember: 31% of Europeans indicate that they cannot recall them.

Table 1. Europeans' 'don't know' rates when asked about the official figures of growth, inflation and unemployment (in percent).

	Growth rate	Inflation rate	Unemployment rate
	% of 'don't know' answers		
European Union	31	31	20
Austria	12	9	6
Belgium	46	44	29
Bulgaria	55	52	41
Croatia	29	30	23
Cyprus	50	52	25
Czech Republic	24	22	11
Denmark	27	24	13
Estonia	15	19	14
Finland	24	24	12
France	37	37	20
Germany	30	27	18
Greece	37	39	11
Hungary	28	22	15
Ireland	23	23	18
Italy	30	30	20
Latvia	52	48	36
Lithuania	40	43	27
Luxembourg	42	41	21
Malta	37	39	35
Netherlands	10	13	9
Poland	13	11	8
Portugal	56	56	37
Romania	67	67	58
Slovakia	44	40	18
Slovenia	32	35	22
Spain	40	42	24
Sweden	12	12	3
United Kingdom	24	23	23

Source: Own elaboration using data from the European Commission (2015a).

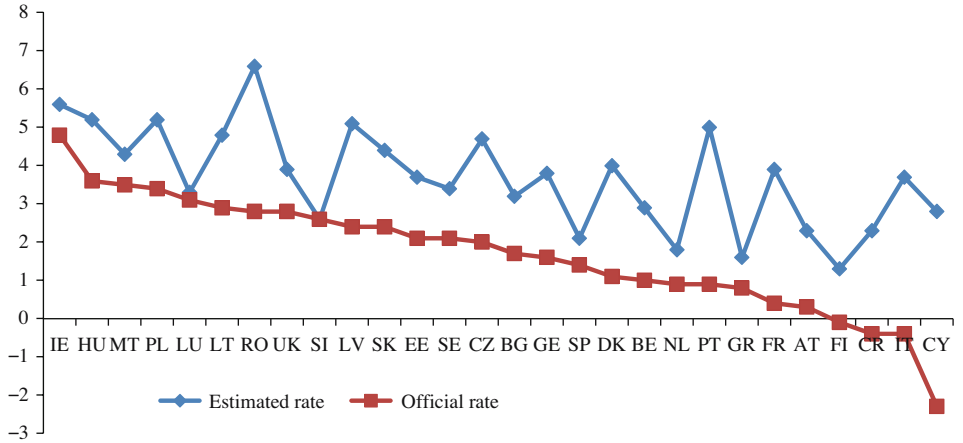

Fig. 1. Estimated and official annual growth rates in the Member States of the European Union in 2014 (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: Countries have been sorted in descending order of the official figure. Countries' acronyms are the following: AT-Austria, BE-Belgium, BG-Bulgaria, CR-Croatia, CY-Cyprus, CZ-Czech Republic, DK-Denmark, EE-Estonia, FI-Finland, FR-France, GE-Germany, GR-Greece, HU-Hungary, IE-Ireland, IT-Italy, LT-Lithuania, LV-Latvia, LU-Luxembourg, MT-Malta, NL-Netherlands, PL-Poland, PT-Portugal, RO-Romania, SE-Sweden, SI-Slovenia, SK-Slovakia, SP-Spain and UK-United Kingdom.

In the case of unemployment, the percentage of 'don't know' answers is 20%. Nonetheless, there is great variation across countries.

Figures 1–3 show the average estimated rates from respondents' answers together with the official rates as reported by the European Commission in its Spring Economic Forecast (2015b,c). These Commission forecasts were released on May 5, 2015. We can see that official numbers are being overestimated practically everywhere.

Concerning growth, figures are overrated in all the countries except for Slovenia and Luxembourg, where respondents provide figures that are very close to the official rates.

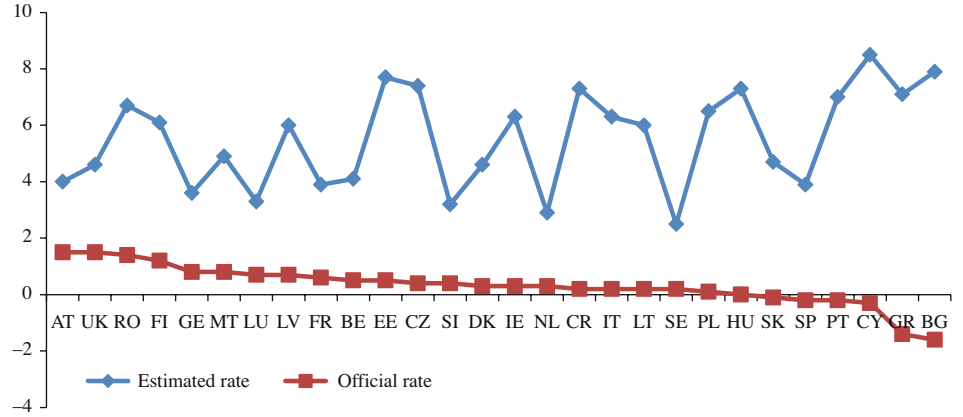

Fig. 2. Estimated and official annual inflation rates in the Member States of the European Union in 2014 (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: See notes under Figure 1.

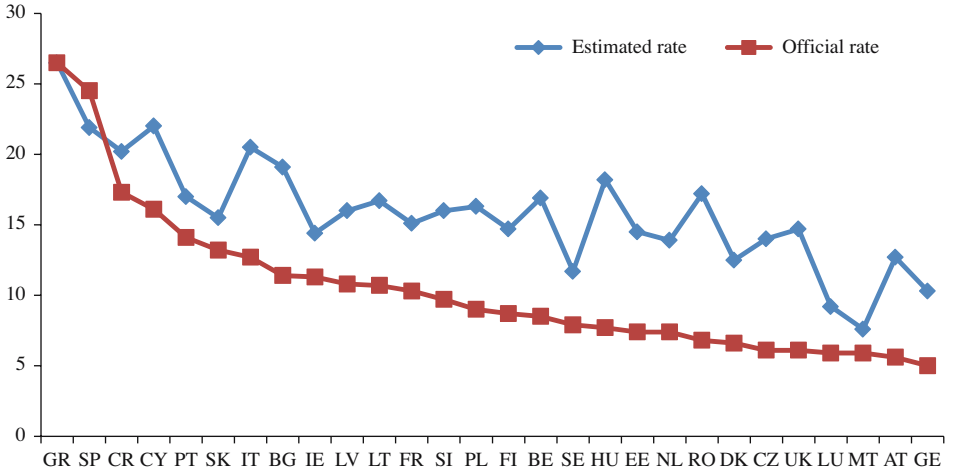

Fig. 3. Estimated and official unemployment rates in the Member States of the European Union in 2014 (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: See notes under Figure 1.

The biggest differences are found in Romania, Portugal, Italy, and Cyprus. In fact, Italy and Cyprus both have negative rates of growth (−0.4% and −2.3%, respectively), whereas the average estimated rates of respondents' answers are positive (3.7% and 2.8%, respectively).

Inflation rates are also overestimated across the Union: while the official rates vary between −1.6% and 1.5%, the estimated rates are between 2.3% and 8.5%. The largest differences are generally observed in countries with negative rates of inflation: in Bulgaria, the average of people's answers is 9.5 points higher than the official rate (7.9% versus −1.6%); in Cyprus the gap is 8.8 points (8.5% versus −0.3%).

Likewise, unemployment figures suffer from overestimation. The only exceptions are Greece and Spain, the countries with the largest figures on unemployment (over 20%): Greek respondents provide a fairly accurate estimate of the official figure; while the Spanish underestimate the rate by 2.6 points.

It is worth noticing that some of the observed overestimation could be due to the guideline values included in the questionnaire. As previously explained, the survey does not contain the appropriate information to check this. Nonetheless, previous research has also found that people tend to overestimate the rates of growth, inflation and unemployment (Papacostas 2008) even when respondents are not provided with any guidelines for answers (Blendon et al. 1997; Curtin 2008, 2009; Malgarini 2009; Giovannini et al. 2015).

Some authors have attempted to provide some explanations for the overestimation of economic figures (European Commission 2007; Curtin 2008; Malgarini 2009). According to Curtin (2008), the overestimation of inflation and unemployment rates is related to a psychological process by which people hold pessimistic views to protect themselves from unexpected events. The European Commission (2007) considers that the overestimation of unemployment figures may well be an indication of the importance that citizens place on this issue. For the particular case of inflation, Malgarini (2009) explains that people might not know the exact meaning of this economic concept as measured by statistical offices.

Table 2. Estimated rates of growth, inflation and unemployment according to respondents' views on the economic situation. European estimated average rates and number of respondents.

Perception of the economic situation	Growth rate	Inflation rate	Unemployment rate
Very good	3.8	3.5	10.0
	(749)	(775)	(828)
Rather good	4.0	4.3	12.9
	(7,000)	(7,032)	(7,746)
Rather bad	3.7	5.1	17.1
	(7,153)	(7,237)	(8,596)
Very bad	3.2	6.2	20.7
	(3,319)	(3,329)	(4,254)
Don't know	4.7	5.6	15.6
	(352)	(356)	(413)
Total	3.7	4.9	15.9
	(18,573)	(18,729)	(21,837)

Source: Own elaboration using data from the European Commission (2015a).
Note: The number of respondents is shown in brackets.

Hence, when they are asked to report a figure, the number they recall is based on their personal experiences. Moreover, people's perceptions are related to their socioeconomic background (i.e., the same increase in prices is felt differently by low-income and high-income households). Additionally, responses might be influenced by people's perceptions on the general economic situation. Since our data have some information on people's opinions about the situation of the national economy, we have tried to analyze whether the overestimation of economic figures can be related to these views. Table 2 reports the European averages estimated from respondents' answers disaggregated according to their views on the economic situation. Results show that most respondents have negative opinions on the situation of the economy. Moreover, we can see that the less favorable opinions on the economy, the higher the reported inflation and unemployment and the lower the growth. Hence, it seems that the overestimation of rates might be related, at least to some extent, to the prevalence of negative views on the economic situation.

Table 3 shows respondents' answers when asked about the evolution of inflation between 2013 and 2014. The percentage of 'don't know' answers is much lower in this case than when people were requested to provide a figure (18% vs. 31%), since it is easier to remember the general evolution of a magnitude than its exact figure. In half of the countries, the most popular answer was that the inflation rate was higher; however, the inflation rate was lower in 2014 than in 2013 in all European Union's countries, except for Latvia.

4. Empirical Approach and Variables

4.1. Empirical Approach

In order to analyze Europeans' knowledge of official economic statistics, we have identified the case in which respondents were asked to provide a figure from the case in which they reported the evolution of the inflation rate.

Table 3. Estimated evolution of the inflation rate between 2013 and 2014 (in percent).

	Higher	Lower	Equal	Don't know	Total
European Union	29	**30**	23	18	100
Austria	50	**25**	19	6	100
Belgium	32	**32**	18	18	100
Bulgaria	34	**14**	21	31	100
Croatia	40	**19**	30	11	100
Cyprus	40	**15**	21	24	100
Czech Republic	27	**32**	31	10	100
Denmark	36	**27**	28	9	100
Estonia	36	**26**	19	19	100
Finland	27	**28**	29	16	100
France	34	**24**	18	24	100
Germany	24	**36**	23	17	100
Greece	40	**19**	24	17	100
Hungary	27	**45**	22	6	100
Ireland	41	**22**	24	13	100
Italy	26	**27**	31	16	100
Latvia	**33**	28	23	16	100
Lithuania	34	**33**	17	16	100
Luxembourg	30	**31**	20	19	100
Malta	28	**29**	19	24	100
Netherlands	36	**40**	19	5	100
Poland	22	**28**	30	20	100
Portugal	34	**19**	23	24	100
Romania	29	**33**	18	20	100
Slovakia	20	**30**	35	15	100
Slovenia	22	**35**	27	16	100
Spain	29	**25**	19	27	100
Sweden	27	**47**	23	3	100
United Kingdom	31	**37**	16	16	100

Source: Own elaboration using data from the European Commission (2015a,b).
Note: Bold numbers indicate the actual evolution of the inflation rate.

In the first case, a two-stage decision process was considered: first, an individual decides whether or not to answer the questions about economic statistics; then, if he decides to answer, he provides a figure.

The most appropriate framework for analyzing an individual's first decision is discrete choice modelling. Since there might be some relationship between those who decide to answer the three questions, a trivariate probit model was considered rather than estimating three separate equations. The model can be defined as a latent variable model as follows (Cappellari and Jenkins 2003):

$$y_m^* = \boldsymbol{\beta}'_{\mathbf{m}}\mathbf{X} + e_m \quad y_m = 1\left[y_m^* > 0\right] \quad m = 1, 2, 3 \tag{1}$$

where y_m^* is a latent variable and only y_m is observed, being a binary variable that takes value 1 if the individual decides to answer question m (0 otherwise); $\mathbf{X}$ is the vector of explanatory variables, and e_m is the error terms that are distributed as multivariate normal

with zero-mean and a variance-covariance matrix with ones in the leading diagonal and correlations $\rho_{jk} = \rho_{kj}$ as off-diagonal elements. The trivariate probit model allows the specification of different sets of regressors for each Equation ($\mathbf{X_m}$). However, in our specification we are interested in checking the influence of the same set of explanatory variables.

In the second stage, for those individuals who decide to answer and provide an estimate, the errors are evaluated. Given that the sample is restricted to those individuals who actually provided a figure, there could be some sample selection bias. The estimation of Heckman selection models indicates that sample selection bias is not an issue here. Hence, we calculate how much their estimates differ (in absolute value) from the corresponding official figure. Following the approach of the first stage, three linear regressions are estimated jointly. The trivariate regression model is specified as follows:

$$z_m = \boldsymbol{\gamma}'_{\mathbf{m}} \mathbf{X} + u_m \quad m = 1, 2, 3 \tag{2}$$

where z_m refers to the magnitude of the errors (in absolute value) made when reporting the figures of GDP, inflation and unemployment, respectively; $\mathbf{X}$ is the vector of explanatory variables, and u_m the error terms that can be correlated between equations. Multivariate regression requires the set of explanatory variables to be the same across all the equations. Moreover, it will produce the same coefficients and standard errors as the ones obtained if the equations are estimated independently. The difference is that the former model allows for obtaining the between-equation covariances.

Since GDP, unemployment, and inflation are measured on different scales, it could be appropriate to consider standardized measures of the absolute errors as dependent variables. Results are shown in the Appendix.

Regarding the evolution of the inflation rate, a multinomial probit model has been specified in order to consider whether respondents gave a right answer, a wrong one or did not know what happened with inflation between 2013 and 2014. The multinomial probit model can be written as:

$$w_j^* = \boldsymbol{\varphi}'_{\mathbf{j}} \mathbf{X} + v_j \quad j = 1, 2, 3 \tag{3}$$

where w_j^* is the latent variable associated with choice j and the observable dependent variable w equals j only if $w_j^* > w_m^*$ for all $j \neq m$; $\mathbf{X}$ is the vector of explanatory variables, and v_j is the error terms which follow independently and identically standard normal distribution. The multinomial probit model has been preferred over multinomial logit model, since it relaxes the independence from irrelevant alternatives (IIA) assumed by the logit (Greene 2011).

4.2. Variables

Europeans' knowledge of economic statistics is measured by a series of variables derived from respondents' answers to the questions included in the survey (Section 3). In the first instance, three dummy variables are considered to take into account whether a respondent agreed to answer the questions on the economic figures. Then, for those who decide to answer, the difference between their estimates and the corresponding official rate is

measured in absolute value. Finally, a categorical variable measures individuals' knowledge of the evolution of inflation. In particular, it is considered whether an individual gave a correct answer, was wrong or did not know.

Explanatory variables have been chosen according to the literature review in Section 2. In particular, Blinder and Krueger (2004) specify that knowledge is a function of an individual's self-interest in the issue, the sources he has checked and his personal characteristics. We complement this approach by taking into account the importance the individual places on official statistics and the individual's trust in them.

Table 4 shows the description of the variables. In the first instance, a measure of an individual's self-interest in economic issues has been defined. This factor is proxied by a dummy variable that indicates whether an individual considers his country's economic situation to be an important issue. In the second instance, daily use of the Internet is included as a proxy for information sources. The Internet has substantially facilitated access to economic information by decreasing associated costs. While television and newspapers seem to be the most common sources of information about the economy (Curtin 2008, 2009; Giovannini et al. 2015), more and more individuals use the Internet to read and watch news. Eurostat (2016e) reports that 45% of the European population read news online in 2012; this figure exceeds 70% in Estonia and Denmark, and exceeds 80% in Finland and Sweden. Moreover, the Internet ranks second (television is ranked first) as the main source of information on national political matters (European Commission 2015d). In this sense, Giovannini et al. (2015) find that together with opinion and political leaders, the Internet is the most significant source of economic information among Italian consumers. Likewise, Curtin (2008) finds evidence of positive correlation between Internet use and knowledge of economic indicators. Nonetheless, Internet use might be linked to some costs for individuals who do not have the ability to find, process and understand information. We control for this fact by including educational attainment. Other socioeconomic features are also considered: gender, age, employment status, social class, and town size. In particular, age, education and income are expected to be positively related to economic knowledge. Compared to younger people, older individuals will *a priori* have better knowledge because they have had more time to gain an understanding of how the economy works and, consequently, the importance of economic figures (Walstad 1997). Nonetheless, the relationship between age and knowledge might be non-linear. People with higher education will also have better knowledge because they are more literate and hence, they can better appreciate and understand the importance and meaning of economic figures. Likewise, people with high income are more likely to be interested and know about economic issues than individuals with low income (Walstad 1997). In addition, we consider two dummy variables that take into account whether an individual trusts official statistics, and whether he considers that political decisions are based on them. Finally, country dummies will be included in the estimations in order to control for cross-country differences.

5. Results

Tables 5, 6, and 8 show the results of the estimations. Table 5 presents the estimates of the trivariate probit regression, which refers to an individual's decision to answer the

Table 4. Description of variables.

Dependent Variables	*Description*
Answer_GDP	Dummy variable that takes value 1 if an individual answered the question about the growth rate of GDP (0 otherwise)
Answer_inflation	Dummy variable that takes value 1 if an individual answered the question about the inflation rate (0 otherwise)
Answer_unemployment	Dummy variable that takes value 1 if an individual answered the question about the unemployment rate (0 otherwise)
E_GDP	Difference between an individual's estimate of the growth rate of GDP in 2014 and the official rate (in absolute value)
E_inflation	Difference between an individual's estimate of the inflation rate in 2014 and the official rate (in absolute value)
E_unemployment	Difference between an individual's estimate of the unemployment rate in 2014 and the official rate (in absolute value)
Inflation_evolution	Categorical variable with the following categories: 1, if an individual gave the right answer regarding the evolution of the inflation rate between 2013 and 2014; 2, if his answer was wrong; 3, if he said he did not know
Independent Variables	*Description*
Economic situation of the country	Dummy variable that takes value 1 if an individual considers that his country's economic situation is an important issue (0 otherwise)
Woman	Dummy variable that takes value 1 if an individual is a woman (0 otherwise)
Age	Individual's age
Education: 16–19	Dummy variable that takes value 1 if an individual stopped his full-time education when he was between 16 and 19 years old
Education: 20+	Dummy variable that takes value 1 if an individual stopped his full-time education when he was at least 20 years old
Education: Still studying	Dummy variable that takes value 1 if an individual is still in full-time education
Education: Don't know/don't answer	Dummy variable that takes value 1 if an individual did not answer the question about his educational attainment

Table 4. Continued.

Independent Variables	*Description*
Area: Small/middle town	Dummy variable that takes value 1 if an individual lives in a small to middle town
Area: Large town	Dummy variable that takes value 1 if an individual lives in a large town
Work: Unemployed	Dummy variable that takes value 1 if an individual is unemployed
Work: Inactive	Dummy variable that takes value 1 if an individual is inactive, that is, he is out of the labor market
Class: Lower middle class	Dummy variable that takes value 1 if an individual considers himself to belong to the lower middle class of the society
Class: Middle class	Dummy variable that takes value 1 if an individual considers himself to belong to the middle class of the society
Class: Upper middle class	Dummy variable that takes value 1 if an individual considers himself to belong to the upper middle class of the society
Class: Higher class	Dummy variable that takes value 1 if an individual considers himself to belong to the higher class of the society
Class: Other	Dummy variable that takes value 1 if an individual considers himself to belong to other social class
Trust	Dummy variable that takes value 1 if an individual tends to trust official statistics
Political_decisions	Dummy variable that takes value 1 if an individual thinks that political decisions are certainly made on the basis of statistical information
Internet use	Dummy variable that takes value 1 if an individual uses the Internet almost everyday
Country	Categorical variable that takes into account the country of residence

Table 5. Trivariate probit regression: decision to answer the questions about growth, inflation and unemployment rates. Estimated coefficients and standard errors.

Variables	Answer_GDP	Answer_inflation	Answer_unemployment
Important issue = Country's economic situation	0.022	0.018	0.050*
	(0.030)	(0.031)	(0.026)
Internet	0.182***	0.174***	0.171***
	(0.027)	(0.026)	(0.025)
Woman	−0.298***	−0.286***	−0.258***
	(0.025)	(0.026)	(0.026)
Age	0.028***	0.029***	0.029***
	(0.004)	(0.003)	(0.004)
Age^2	−0.000***	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)
Education = 16–19	0.064**	0.058**	0.112***
	(0.027)	(0.029)	(0.033)
Education = 20+	0.231***	0.200***	0.245***
	(0.035)	(0.039)	(0.045)
Education = Still studying	0.171**	0.128*	0.205***
	(0.081)	(0.072)	(0.075)
Education = Don't know/don't answer	−0.091	−0.138*	−0.208***
	(0.075)	(0.080)	(0.062)
Area = Small/middle town	0.016	−0.017	0.020
	(0.050)	(0.050)	(0.044)
Area = Large town	0.049	−0.028	−0.018
	(0.049)	(0.055)	(0.044)
Work = Unemployed	−0.071*	−0.112***	−0.044
	(0.037)	(0.041)	(0.038)
Work = Inactive	−0.030	0.007	−0.037*
	(0.026)	(0.024)	(0.022)

Table 5. Continued.

Variables	Answer_GDP	Answer_inflation	Answer_unemployment
Class = Lower middle class	0.163***	0.139***	0.137***
	(0.040)	(0.042)	(0.037)
Class = Middle class	0.152***	0.132***	0.124***
	(0.036)	(0.034)	(0.030)
Class = Upper middle class	0.331***	0.297***	0.235***
	(0.054)	(0.062)	(0.059)
Class = Higher class	0.357***	0.509***	0.293*
	(0.133)	(0.139)	(0.156)
Class = Other	−0.194***	−0.223***	−0.315***
	(0.063)	(0.061)	(0.074)
Trust	0.169***	0.157***	0.137***
	(0.022)	(0.022)	(0.023)
Political_decisions	0.117***	0.119***	0.109***
	(0.037)	(0.036)	(0.039)
Constant	−0.585***	−0.524***	−0.023
	(0.109)	(0.087)	(0.095)

Note: Reference categories are the following: men, with education up to 15 years old, living in rural areas or villages, employed, working class, individuals who do not use the Internet, those who do not consider their country's economic situation to be an important issue, who tend not to trust official statistics and who do not think that political decisions are made on the basis of statistical information. Standard errors are shown in brackets. As usual ***, ** and * indicate statistically significant figures at the 1, 5 and 10 percent levels, respectively.

Table 5. Trivariate probit regression: decision to answer the questions about growth, inflation and unemployment rates. Estimated coefficients and standard errors (continued).

Countries	Answer_ GDP	Answer_ inflation	Answer_ unemployment	Countries	Answer_ GDP	Answer_ inflation	Answer_ unemployment
Austria	0.894***	0.953***	0.688***	Latvia	−0.395***	−0.283***	−0.495***
	(0.017)	(0.019)	(0.017)		(0.009)	(0.011)	(0.011)
Belgium	−0.157***	−0.138***	−0.308***	Lithuania	−0.087***	−0.136***	−0.238***
	(0.017)	(0.016)	(0.015)		(0.007)	(0.009)	(0.009)
Bulgaria	−0.424***	−0.328***	−0.545***	Luxembourg	−0.280***	−0.258***	−0.205***
	(0.018)	(0.020)	(0.018)		(0.016)	(0.015)	(0.014)
Croatia	0.216***	0.156***	−0.092***	Malta	−0.060***	−0.099***	−0.501***
	(0.020)	(0.020)	(0.019)		(0.022)	(0.024)	(0.020)
Cyprus	−0.335***	−0.391***	−0.144***	Netherlands	0.634***	0.470***	0.183***
	(0.021)	(0.022)	(0.020)		(0.020)	(0.021)	(0.020)
Czech Rep.	0.373***	0.462***	0.367***	Poland	0.821***	0.891***	0.594***
	(0.014)	(0.016)	(0.013)		(0.015)	(0.015)	(0.012)
Denmark	0.028**	0.127***	0.028*	Portugal	−0.393***	−0.385***	−0.374***
	(0.014)	(0.017)	(0.017)		(0.021)	(0.022)	(0.018)
Estonia	0.677***	0.451***	0.158***	Romania	−0.763***	−0.755***	−1.044***
	(0.013)	(0.014)	(0.013)		(0.023)	(0.023)	(0.023)
Finland	0.299***	0.303***	0.232***	Slovakia	−0.177***	−0.116***	0.032**
	(0.011)	(0.010)	(0.012)		(0.016)	(0.016)	(0.014)
Germany	0.212***	0.275***	0.086***	Slovenia	0.150***	0.056***	−0.068***
	(0.009)	(0.010)	(0.009)		(0.012)	(0.013)	(0.011)
Greece	0.008	−0.057**	0.376***	Spain	0.039**	−0.052***	−0.016
	(0.021)	(0.022)	(0.021)		(0.016)	(0.017)	(0.014)
Hungary	0.286***	0.474***	0.184***	Sweden	0.577***	0.660***	0.696***
	(0.016)	(0.018)	(0.014)		(0.013)	(0.016)	(0.016)

Table 5. Continued.

Countries	Answer_ GDP	Answer_ inflation	Answer_ unemployment	Countries	Answer_ GDP	Answer_ inflation	Answer_ unemployment
Ireland	0.413***	0.414***	0.118***	United Kingdom	0.316***	0.376***	−0.163***
	(0.014)	(0.016)	(0.015)		(0.016)	(0.015)	(0.016)
Italy	0.170***	0.127***	−0.023*				
	(0.012)	(0.012)	(0.013)				
rho21	1.377***	(0.029)					
rho31	1.076***	(0.035)					
rho32	1.151***	(0.036)					
Likelihood ratio test of rho21 = rho31 = rho32 = 0			chi2(3) = 22579.5 Prob > chi2 = 0.0000				

Note: France is the reference country. Standard errors are shown in brackets. As usual ***, ** and * indicate statistically significant figures at the 1, 5 and 10 percent levels, respectively.

Table 6. Joint regression on the errors made when reporting the figures of growth, inflation and unemployment. Estimated coefficients and standard errors.

Variables	E_GDP	E_inflation	E_unemployment
Important issue = Country's economic situation	−0.336***	−0.370***	−0.241**
	(0.055)	(0.087)	(0.108)
Internet	−0.433***	−0.864***	−0.729***
	(0.062)	(0.098)	(0.122)
Woman	0.579***	0.855***	1.290***
	(0.046)	(0.073)	(0.090)
Age	−0.056***	−0.082***	−0.137***
	(0.009)	(0.014)	(0.017)
Age^2	0.000***	0.000***	0.001***
	(0.000)	(0.000)	(0.000)
Education = 16–19	−0.224***	−0.332***	−0.688***
	(0.080)	(0.126)	(0.157)
Education = 20+	−0.709***	−0.988***	−1.255***
	(0.087)	(0.137)	(0.170)
Education = Still studying	−0.272*	−0.786***	−1.847***
	(0.153)	(0.241)	(0.300)
Education = Don't know/don't answer	−0.425**	−0.439	−0.874**
	(0.176)	(0.278)	(0.346)
Area = Small/middle town	−0.073	−0.195**	−0.291***
	(0.056)	(0.089)	(0.110)
Area = Large town	−0.219***	−0.382***	−0.403***
	(0.062)	(0.098)	(0.122)
Work = Unemployed	0.156*	0.275*	0.481***
	(0.089)	(0.141)	(0.176)
Work = Inactive	−0.086	0.169	0.174
	(0.071)	(0.112)	(0.140)

Table 6. Continued.

Variables	E_GDP	E_inflation	E_unemployment
Class = Lower middle class	−0.375***	−0.643***	−1.272***
	(0.073)	(0.115)	(0.143)
Class = Middle class	−0.203***	−0.628***	−0.948***
	(0.062)	(0.098)	(0.122)
Class = Upper middle class	−0.414***	−0.936***	−1.456***
	(0.100)	(0.158)	(0.196)
Class = Higher class	−0.085	−0.398	−0.641
	(0.244)	(0.386)	(0.480)
Class = Other	−0.265*	−0.925***	−1.120***
	(0.149)	(0.235)	(0.292)
Trust	−0.076	−0.506***	−1.008***
	(0.047)	(0.075)	(0.093)
Political_decisions	−0.219***	−0.167	−0.469***
	(0.065)	(0.102)	(0.127)
Constant	6.747***	8.818***	12.665***
	(0.263)	(0.416)	(0.517)

Note: See notes under Table 5.

Table 6. Joint regression on the errors made when reporting the figures of growth, inflation and unemployment. Estimated coefficients and standard errors (continued).

Countries	E_GDP	E_inflation	E_unemployment	Countries	E_GDP	E_inflation	E_unemployment
Austria	−1.040***	−0.825***	1.198***	Latvia	−0.267	1.970***	0.462
	(0.159)	(0.252)	(0.313)		(0.190)	(0.301)	(0.374)
Belgium	−1.353***	0.057	10.045***	Lithuania	−0.991***	2.134***	1.355***
	(0.179)	(0.283)	(0.352)		(0.181)	(0.285)	(0.355)
Bulgaria	−0.593***	5.352***	2.210***	Luxembourg	−1.421***	−0.913**	−1.791***
	(0.191)	(0.302)	(0.376)		(0.225)	(0.355)	(0.441)
Croatia	−0.162	3.075***	0.057	Malta	−1.212***	0.292	−1.288***
	(0.172)	(0.271)	(0.337)		(0.220)	(0.348)	(0.432)
Cyprus	1.853***	5.036***	1.343***	Netherlands	−2.188***	−0.717***	1.759***
	(0.238)	(0.376)	(0.468)		(0.162)	(0.256)	(0.319)
Czech Rep.	−0.816***	2.625***	5.969***	Poland	−1.175***	2.179***	2.499***
	(0.166)	(0.263)	(0.326)		(0.162)	(0.256)	(0.318)
Denmark	−0.405**	0.972***	1.322***	Portugal	0.396**	2.999***	−0.841**
	(0.169)	(0.267)	(0.332)		(0.199)	(0.314)	(0.390)
Estonia	−1.374***	3.127***	2.415***	Romania	0.137	1.724***	3.708***
	(0.165)	(0.260)	(0.323)		(0.222)	(0.351)	(0.436)
Finland	−0.843***	1.618***	0.816**	Slovakia	−1.433***	0.730**	−2.355***
	(0.166)	(0.263)	(0.326)		(0.181)	(0.286)	(0.355)
Germany	−1.721***	−1.061***	−0.381	Slovenia	−1.918***	−1.048***	0.959***
	(0.155)	(0.245)	(0.304)		(0.175)	(0.276)	(0.343)
Greece	−0.593***	4.930***	−2.542***	Spain	−1.948***	0.081	−1.080***
	(0.179)	(0.283)	(0.352)		(0.177)	(0.279)	(0.347)
Hungary	−1.255***	3.195***	4.033***	Sweden	−1.407***	−0.620**	−1.220***
	(0.167)	(0.265)	(0.329)		(0.160)	(0.253)	(0.315)
Ireland	−1.077***	1.809***	−0.175	United Kingdom	−1.235***	−0.403	2.784***
	(0.165)	(0.261)	(0.324)		(0.162)	(0.256)	(0.318)
Italy	0.519***	2.491***	3.274***				
	(0.170)	(0.268)	(0.333)				

Breusch-Pagan test of independence: chi2(3) = 4808.98, Pr = 0.0000

Note: See notes under Table 5.

questions about the figures of GDP growth, inflation and unemployment. Results show that most of the explanatory variables are statistically significant and perform as expected. The only exceptions are those related to the perceived importance of the economic situation and the type of area. In particular, the fact that a national economic situation is considered to be important only has a significant effect (at the ten percent level) on the probability of answering the question on unemployment, but not for GDP or inflation; meanwhile, the type of area is not statistically significant in any of the three equations. In contrast, daily Internet use has a significant positive effect on the probability of answering any of the questions. The positive coefficients indicate that people using the Internet are more likely to answer these questions than those who do not use it. As regards individuals' personal characteristics, our estimates confirm the existence of wide variation among socio-economic groups. In particular, we observe that groups that have been traditionally considered as socially disadvantaged are less likely to answer. We find, for instance, that unemployed individuals are less likely to answer than those in work. This likelihood increases with the level of education and social class. Note how the magnitudes of the coefficients of education and class increase as their levels rise. Hence, the higher the education and social class, the more likely they are to answer. While there is an indication that the age relationship is non-linear, the non-linear impact is not strong across the ages in the survey; older people are more likely to answer the questions. Moreover, country dummies are statistically significant, mostly at the one percent level. While the interpretation of all those estimates is quite complex, finding them significant indicates the existence of important cross-country differences with regard to the economic knowledge of their respective populations. As previously mentioned, these differences might be related to some extent to the effect of providing common guidelines for answers. The p-value of the likelihood ratio test leads to the rejection of the null hypothesis of the absence of correlation between the equations, which suggests the suitability of the proposed trivariate probit compared to the estimation of three independent equations.

Table 6 presents the estimates for the trivariate regressions of the errors made when reporting the rates of GDP growth, inflation and unemployment. Negative coefficients indicate smaller errors and subsequently better knowledge of economic indicators; correspondingly, positive coefficients suggest larger errors and thus poorer knowledge. In this case, all the considered determinants are statistically significant. The appropriateness of the joint estimation has been checked using the Breusch–Pagan test.

Results show that people who regard the economic situation as important and who use the Internet are less likely to make large errors than those who do not. Likewise, individuals who declare that they trust official statistics and believe that political decisions are based on them tend to make smaller errors. As previous research has noted, the value of official statistics is directly linked to people's trust and their perceived usefulness (Gore et al. 1991; Giovanni et al. 2008, 2015; Holt 2008; Papacostas 2008). The pattern of variation across socioeconomic groups is quite similar to the one observed when analyzing the decision to answer, and it broadly confirms the main findings of the literature. Hence, the largest errors and, consequently, the worst knowledge, are observed among young and socially disadvantaged groups (Walstad 1997; Walstad and Rebeck 2002; Blinder and Krueger 2004; Curtin 2008, 2009; Malgarini 2009; Giovannini et al. 2015). In particular, respondents with education up to 15 years old, who belong to the working class, who do

not work or are out of the labor market and live in rural areas, tend to make the largest errors. Our estimates also show that women tend to make larger errors than men. This result confirms previous findings in the literature (Bryan and Venkatu 2001; Walstad and Rebeck 2002; Blinder and Krueger 2004; Curtin 2008, 2009; Giovannini et al. 2015). Such gender gaps might be related to different levels of economic/financial literacy (Bruine de Bruin et al. 2010; Burke and Manz 2014); in fact, women appear to be much more financially illiterate compared to men (Lusardi and Mitchell 2008; Fonseca et al. 2012). In the particular case of inflation, it has also been suggested that gender differences might be related to men and women's different shopping experiences (Jonung 1981; Jonung and Laidler 1988; Bryan and Venkatu 2001).

Country dummies are again statistically significant. In order to shed some light on these country effects, Figures 4–6 show countries ordered by the size of the errors for growth, inflation and unemployment, respectively. For GDP, the largest errors are found in Cyprus, followed by Italy and Portugal; for inflation, in Bulgaria, Cyprus and Greece; and for unemployment, in Hungary and Romania. Such country effects could be related to both demand and supply side factors. On the demand side, there might be differences between countries in terms of the importance that their populations place on official statistics and whether they trust them. In this sense, it could be conjectured that countries in which citizens trust official statistics and find them important, would have smaller errors. Since our database contains this information at the individual level, these two factors have already been considered in our models. As previously explained, results indicate that those individuals who trust official statistics and believe that they are the base for political decisions make smaller errors. Accordingly, countries in which the levels of trust in official statistics were high and where citizens believed that they were used for policy-making, would tend to have better knowledge of economic figures. Additionally, the observed country effect could be related to differences in a population's levels of economic/

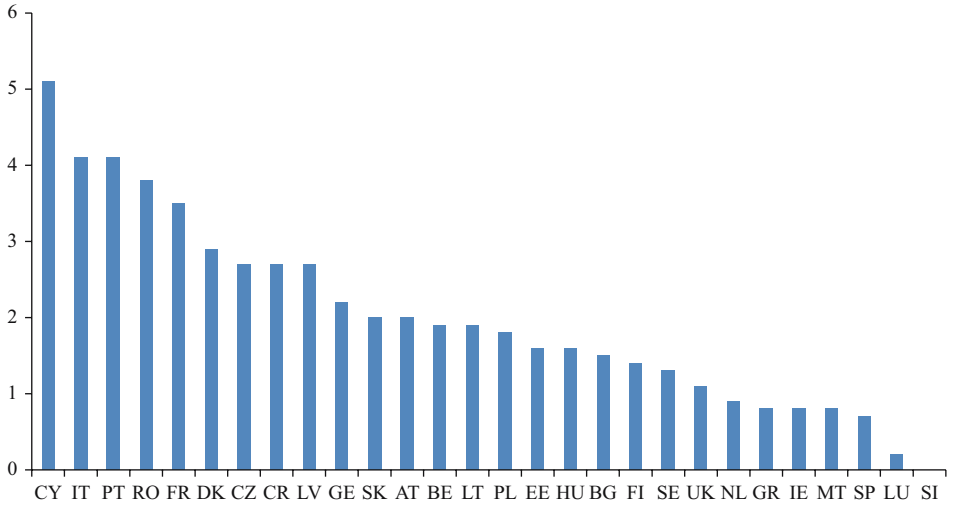

Fig. 4. Errors in growth reporting by country (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: Errors have been calculated as the difference between the average estimated rate and the official rate. For the meaning of the acronyms, see notes under Figure 1.

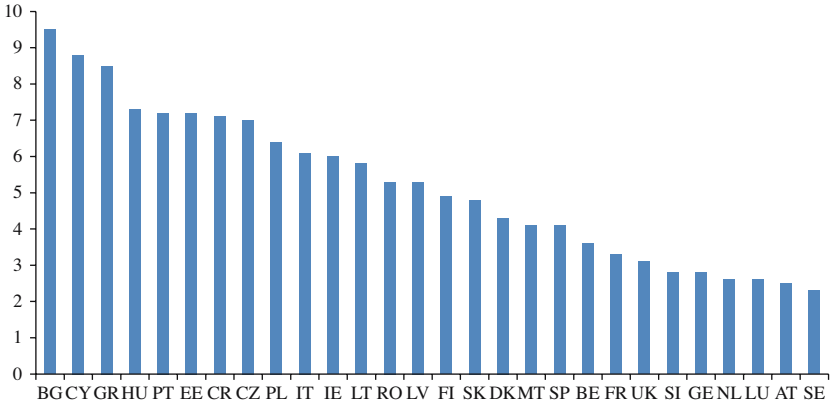

Fig. 5. Errors in inflation reporting by country (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: See notes under Figure 4.

financial literacy. When individuals are economically/financially literate, they are able to choose the relevant information and apply it efficiently in decision-making (Bruine de Bruin et al. 2010; Burke and Manz 2014). Hence, countries with highly economically/financially literate populations would be expected to have lower errors. In the case of the European Union, some recent data indicate that there is great variation in levels of financial literacy across the Member States: from 22% of adults being financially literate in Romania to 71% in Sweden and Denmark (Klapper et al. 2015). The correlation coefficients between the errors and the levels of financial literacy are −0.43 for GDP, −0.55 for inflation and 0 for unemployment. The negative signs for GDP and inflation imply that countries with higher literacy tend to show lower errors. In fact, if we include

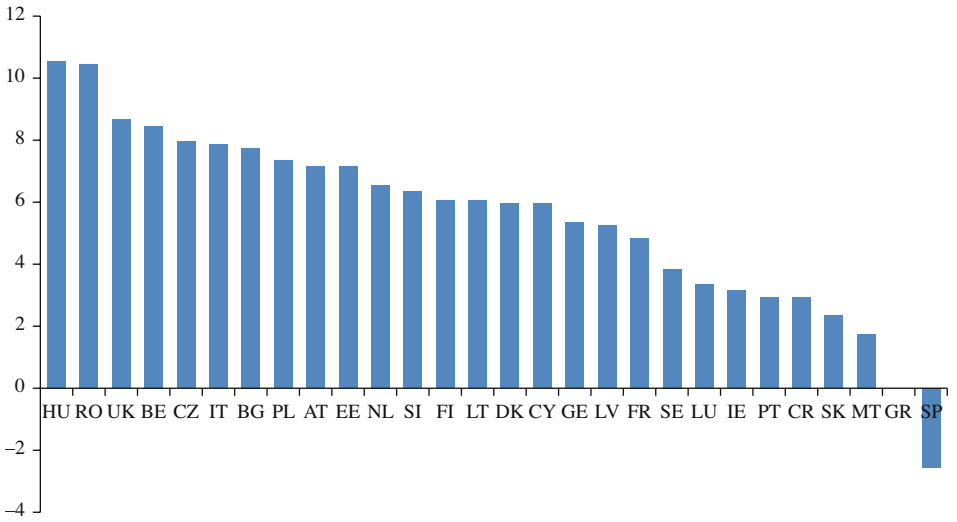

Fig. 6. Errors in unemployment reporting by country (in percent). Source: Own elaboration using data from the European Commission (2015a,c). Notes: See notes under Figure 4. Spain has a negative error due to the underestimation of the official rate of unemployment.

financial literacy as an explanatory variable in the trivariate regression on the errors, we find that it is statistically significant at the 1 percent level in the case of GDP and inflation: the estimated coefficients (and standard errors) are: −0.033 (0.002), −0.080 (0.002) and 0.001 (0.004) for the equations of GDP, inflation and unemployment, respectively. Such results might indicate that the more complexity involved in the concepts and figures of GDP and inflation as compared to those of unemployment.

On the supply side, countries might differ in how data are released and how the media reports these data. In this sense, we could expect that countries in which data were reported more widely and accurately would have lower errors. Curtin (2008, 2009) investigated the coverage of official statistics by the American media and found that there was some inadequate communication of the figures of inflation and growth: they were less frequently reported in the media than unemployment, and they were generally reported in qualitative terms rather than quantitative. Hence, the low level of knowledge of the population might be explained, at least to some extent, by inadequate coverage by the media.

As to data release policies, the Internet has become the main channel for wide dissemination of official figures. National statistical offices publish all new data releases on their websites, where they can be accessed by everyone. Furthermore, online social networks are increasingly used to announce data releases. In this sense, we have studied the presence of statistical agencies on two of the most popular social networks (i.e., Twitter and Facebook). Though data releases published on social networks do not reach the full population (but only those who use the social networks), it is worth studying them, given the increasing importance of social networks. Table 7 summarizes the presence of statistical offices on Twitter.

From January 2017, all European statistical agencies have a Twitter account except for Slovakia and Bulgaria. Luxembourg was first to open an account, in 2007, followed by Estonia, Ireland, the Netherlands, Slovenia, and Sweden in 2009. The last countries to join were Malta and Poland in January and November 2016, respectively. Some statistical offices have accounts in more than one language – usually their national language and English (e.g., Belgium, Germany, and Finland, among others). There is great variation in the use of this network among countries: the United Kingdom is by far the most active user, and also has the largest number of followers. Other countries that stand out are Spain, France, and the Netherlands. Though all institutions use this network to disseminate and announce their data releases, the type of messages published varies: for example, some of them mainly report economic news in qualitative rather than quantitative terms, some include tables with the tweets, and others just post the announcement of a new press release. With regard to Facebook, less than half of the offices have a profile on this social network. Unlike Twitter, which directly provides the figure of the total number of tweets, it is more difficult to know the total number of posts. Moreover, we note that interaction with users varies across agencies: some have implemented the option to send a message while others have not. In addition, the time they take to answer varies: from very responsive to messages to no information. In order to get some understanding on whether errors might be related to dissemination actions on online social networks, we have calculated the correlation coefficient between the errors and the average number of tweets per year, −0.03, −0.36, and −0.12 for GDP, inflation and unemployment, respectively. Interestingly enough, the figures are negative, which suggests that intense activity on

Table 7. Presence of national statistical offices on Twitter.

	Tweets	Followers	Joining date	Tweets per year	English account
Austria	1,027	3,711	May-14	384	
Belgium	1,144	1,039	Feb-12	232	X
Bulgaria	-	-	-	-	
Croatia	1,982	1,545	Dec-13	642	X
Cyprus	692	422	Feb-14	237	
Czech Republic	2,632	8,841	Mar-11	450	
Denmark	2,418	9,867	Nov-13	763	
Estonia	808	1,301	Nov-09	113	
Finland	3,259	5,199	Feb-11	550	X
France	9,180	49,700	Nov-10	1,487	
Germany	3,256	10,900	Feb-11	550	X
Greece	2,455	3,247	Dec-10	403	
Hungary	1,005	66	Dep-14	430	X
Ireland	2,432	14,600	Aug-09	328	X
Italy	6,843	50,900	Sep-10	1,079	X
Latvia	2,726	4,333	Jul-11	495	X
Lithuania	1,936	413	Jan-12	387	
Luxembourg	458	691	Jun-07	48	
Malta	484	322	Jan-16	483	X
Netherlands	13,500	113,000	Jul-09	1,798	X
Poland	195	713	Oct-16	774	X
Portugal	2,409	2,089	Aug-12	545	
Romania	43	292	Jul-15	29	
Slovakia	-	-	-	-	
Slovenia	2,809	2,809	Jul-09	374	X
Spain	8,633	28,800	Aug-11	1,591	
Sweden	5,677	8,401	Feb-09	717	X
United Kingdom	13,300	245,000	Nov-10	2,155	X

Source: Own elaboration. Date of collection: January 12, 2017.

social networks (i.e., a high number of tweets) is associated with low errors. Though this finding is suggestive, it is important to take into account that it does not imply any causation. Moreover, we must note that the values of the coefficients are low.

Finally, Table 8 presents the estimates in the multinomial probit model on the evolution of the inflation rate. As previously indicated, the rate of people answering 'don't know' largely decreased when they were asked about the evolution. Nonetheless, the profile of individuals who answered correctly is very similar to the one observed in the other questions: people who use the Internet, men, with higher education levels or even still studying and in the upper middle or higher classes of society, who trust official statistics and think these are the basis for political decisions are those most likely to answer correctly. In contrast, neither the type of area nor the perceived importance of the economic situation has any significant influence.

6. Summary of Findings and Concluding Remarks

The aim of this article has been to assess Europeans' knowledge of some key economic statistics (i.e., GDP, inflation and unemployment) and to identify the determinants of such knowledge. Results show two main problems: people who don't know, and people who think they know when, in fact, they don't.

On the one hand, the rate of 'don't know' answers are quite important: almost one out of three Europeans indicate that they do not know the national rates of GDP or inflation, and about one out of five cannot report the unemployment rate or the evolution of prices between 2013 and 2014.

On the other hand, the level of knowledge of those who attempted to provide a figure is generally low: respondents' answers differed from the official figures up to five percentage points in the case of the growth rate, and up to ten percentage points for inflation and unemployment rates. We observe that there is a general tendency to overestimate official figures. In some countries, such overestimation suggests a complete misperception of the economic reality: people reporting positive rates of growth when the actual rates are negative (e.g., Cyprus and Italy); prices increasing when they are actually decreasing (e.g., Greece and Bulgaria).

Such gaps of knowledge appear to be shaped not only by individuals' socioeconomic characteristics, but also by their trust in official statistics and their perceived usefulness. Moreover, cross-country differences seem to be related to a population's level of financial literacy, and also possibly to data release policies.

In light of these findings, some recommendations can be suggested in order to improve knowledge of these statistics and to stimulate their demand and use. As Gabriel Tarde (1903) predicted, we have reached a point in time when very accurate statistical information is readily available to the public. However, people don't know this, or think they know when they actually don't; and they do not use statistics properly for decision-making. Even when official statistics are readily available, individuals tend to rely on less accurate sources of information (Cavallo et al. 2016). Hence, it seems crucial that producers of official statistics design campaigns to communicate the value of official statistics to the general public. Such campaigns should aim at making people understand the reasons why official statistics are important, that is, to make them aware of their

Table 8. Multinomial probit regression on the estimated evolution of inflation. Marginal effects.

Variables	Wrong answer	Right answer	Don't know/don't answer
Important issue = Country's economic situation	−0.002	0.007	−0.004
	(0.008)	(0.006)	(0.006)
Internet	−0.007	0.050***	−0.043***
	(0.011)	(0.010)	(0.007)
Woman	0.007	−0.066***	0.059***
	(0.008)	(0.008)	(0.005)
Age	0.001	0.008***	−0.009***
	(0.002)	(0.002)	(0.001)
Age^2	−0.000*	−0.000***	0.000***
	(0.000)	(0.000)	(0.000)
Education = 16–19	−0.012	0.039***	−0.027***
	(0.010)	(0.012)	(0.009)
Education = 20+	−0.013	0.064***	−0.051***
	(0.011)	(0.012)	(0.010)
Education = Still studying	−0.006	0.072***	−0.066***
	(0.024)	(0.025)	(0.014)
Education = Don't know/don't answer	−0.036	0.005	0.030*
	(0.024)	(0.030)	(0.018)
Area = Small/middle town	−0.009	0.004	0.005
	(0.015)	(0.010)	(0.011)
Area = Large town	0.004	0.005	−0.009
	(0.017)	(0.012)	(0.012)
Work = Unemployed	−0.024**	−0.006	0.030***
	(0.011)	(0.010)	(0.009)
Work = Inactive	−0.000	−0.007	0.008
	(0.012)	(0.011)	(0.008)

Table 8. Continued.

Variables	Wrong answer	Right answer	Don't know/don't answer
Class = Lower middle class	−0.020	0.060***	−0.039***
	(0.015)	(0.010)	(0.011)
Class = Middle class	−0.019	0.060***	−0.041***
	(0.013)	(0.010)	(0.007)
Class = Upper middle class	−0.048**	0.108***	−0.060***
	(0.020)	(0.016)	(0.012)
Class = Higher class	−0.042	0.103***	−0.061***
	(0.042)	(0.039)	(0.024)
Class = Other	−0.104***	−0.019	0.122***
	(0.025)	(0.013)	(0.024)
Trust	−0.015	0.049***	−0.035***
	(0.011)	(0.009)	(0.007)
Political_decisions	0.017	0.024**	−0.041***
	(0.015)	(0.012)	(0.011)

Note: See notes under Table 5.

Table 8. Multinomial probit regression on the estimated evolution of inflation. Marginal effects (continued).

Countries	Wrong answer	Right answer	Don't know/don't answer	Countries	Wrong answer	Right answer	Don't know/don't answer
Austria	0.189***	−0.018***	−0.171***	Latvia	−0.006*	0.086***	−0.080***
	(0.004)	(0.003)	(0.004)		(0.004)	(0.003)	(0.003)
Belgium	−0.038***	0.096***	−0.058***	Lithuania	−0.017***	0.078***	−0.061***
	(0.006)	(0.004)	(0.005)		(0.003)	(0.003)	(0.004)
Bulgaria	0.045***	−0.098***	0.052***	Luxembourg	−0.026***	0.033***	−0.007*
	(0.005)	(0.004)	(0.007)		(0.005)	(0.003)	(0.004)
Croatia	0.146***	−0.047***	−0.098***	Malta	−0.041***	0.035***	0.006
	(0.005)	(0.004)	(0.004)		(0.006)	(0.005)	(0.007)
Cyprus	0.093***	−0.090***	−0.003	Netherlands	0.056***	0.085***	−0.141***
	(0.006)	(0.004)	(0.006)		(0.007)	(0.006)	(0.004)
Czech Rep.	0.052***	0.078***	−0.130***	Poland	0.000	0.032***	−0.032***
	(0.004)	(0.004)	(0.003)		(0.004)	(0.003)	(0.004)
Denmark	0.130***	−0.031***	−0.099***	Portugal	0.057***	−0.015***	−0.041***
	(0.006)	(0.004)	(0.003)		(0.006)	(0.006)	(0.007)
Estonia	0.029***	−0.015***	−0.014***	Romania	−0.045***	0.083***	−0.038***
	(0.004)	(0.003)	(0.004)		(0.006)	(0.005)	(0.005)
Finland	0.042***	0.009**	−0.051***	Slovakia	0.036***	0.049***	−0.085***
	(0.005)	(0.004)	(0.004)		(0.005)	(0.004)	(0.004)
Germany	−0.029***	0.099***	−0.069***	Slovenia	−0.036***	0.114***	−0.078***
	(0.003)	(0.002)	(0.003)		(0.003)	(0.003)	(0.003)
Greece	0.142***	−0.057***	−0.086***	Spain	−0.044***	0.031***	0.013**
	(0.006)	(0.004)	(0.006)		(0.005)	(0.005)	(0.005)
Hungary	−0.050***	0.226***	−0.176***	Sweden	−0.007	0.193***	−0.187***
	(0.004)	(0.004)	(0.003)		(0.007)	(0.007)	(0.003)
Ireland	0.124***	−0.035***	−0.090***	United Kingdom	−0.105***	0.183***	−0.078***
	(0.004)	(0.003)	(0.004)		(0.005)	(0.006)	(0.004)
Italy	0.054***	0.030***	−0.084***				
	(0.004)	(0.003)	(0.004)				

Note: See notes under Table 5.

usefulness. They should highlight how this economic information is used in daily life and how the misunderstanding of economic figures could lead to poor decision-making. The misperception of inflation figures may have an impact on household consumption and decisions on savings (Duffy and Lunn 2009; Carrillo and Emran 2012); the misperception of unemployment rates may alter people's economic and political attitudes (Kunovich 2012) and labor outcomes (Cardoso et al. 2016). For instance, people, who overestimate unemployment figures may consider that they have little bargaining power and hence lower their reservation wages, that is, the lowest wage at which they are willing to work, which would thus result in a lower actual wage (Cardoso et al. 2016). Additionally, economic knowledge influences people's opinion on public issues (Blendon et al. 1997; Walstad 1997; Walstad and Rebeck 2002; Blinder and Krueger 2004). Becoming better informed is not only self-serving, but also provides an improved basis for decision-making. As Stigler (1962, 103) indicated in the case of the labor market: "The information a man possesses on the labor market is capital: it was produced at the cost of search, and it yields a higher wage rate than on average would be received in its absence". While the cost of search has dramatically decreased thanks to new technologies, having the most accurate information is crucial for achieving the best possible outcomes. Special efforts should be devoted to communicating GDP and inflation figures since our findings suggest that that these two indicators are the least known by the general public. Communication campaigns should try to clearly explain how these economic figures are produced and the guarantees the production process offers in order to foster people's trust in them. Moreover, and given the low levels of knowledge, programs of statistical literacy for the adult population appears to be a must. While target population for this kind of program is usually students, our results suggest that they should also address the adult population. In particular, findings indicate that the demographic groups most in need of help are those who are least educated, in lower social classes, unemployed and women. Hence, the importance of designing target training actions for these groups in order to improve their understanding and knowledge of official statistics; and also to show them how they can use statistics for decision-making. Additionally, the role of online social networks as a channel for statistical news should be further explored. Our analysis reveals the existence of important differences in the use of these social networks by statistical offices. In fact, some agencies have not started using them. Given the increasing number of users, it is very likely that they will be a major tool for the wide diffusion of statistical news in the near future.

Appendix

Table A1. Joint regression on the absolute errors (z-scores). Estimated coefficients and standard errors.

Variables	z_GDP	z_inflation	z_unemployment
Important issue = Country's economic situation	−0.108***	−0.071***	−0.037**
	(0.018)	(0.017)	(0.016)
Internet	−0.140***	−0.166***	−0.111***
	(0.020)	(0.019)	(0.019)
Woman	0.187***	0.165***	0.196***
	(0.015)	(0.014)	(0.014)
Age	−0.018***	−0.016***	−0.021***
	(0.003)	(0.003)	(0.003)
Age^2	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)
Education = 16–19	−0.072***	−0.064***	−0.104***
	(0.026)	(0.024)	(0.024)
Education = 20+	−0.228***	−0.190***	−0.191***
	(0.028)	(0.026)	(0.026)
Education = Still studying	−0.088*	−0.151***	−0.280***
	(0.049)	(0.046)	(0.046)
Education = Don't know/don't answer	−0.137**	−0.084	−0.133**
	(0.057)	(0.054)	(0.052)
Area = Small/middle town	−0.023	−0.038**	−0.044***
	(0.018)	(0.017)	(0.017)
Area = Large town	−0.071***	−0.074***	−0.061***
	(0.020)	(0.019)	(0.018)
Work = Unemployed	0.050*	0.053*	0.073***
	(0.029)	(0.027)	(0.027)
Work = Inactive	−0.028	0.032	0.026
	(0.023)	(0.022)	(0.021)
Class = Lower middle class	−0.121***	−0.124***	−0.193***
	(0.024)	(0.022)	(0.022)
Class = Middle class	−0.066***	−0.121***	−0.144***
	(0.020)	(0.019)	(0.019)
Class = Upper middle class	−0.134***	−0.180***	−0.221***
	(0.032)	(0.030)	(0.030)
Class = Higher class	−0.027	−0.077	−0.097
	(0.079)	(0.074)	(0.073)
Class = Other	−0.085*	−0.178***	−0.170***
	(0.048)	(0.045)	(0.044)
Trust	−0.025	−0.097***	−0.153***
	(0.015)	(0.014)	(0.014)
Political_decisions	−0.071***	−0.032	−0.071***
	(0.021)	(0.020)	(0.019)
Constant	1.167***	0.671***	0.840***
	(0.085)	(0.080)	(0.079)

Note: See notes under Table 5. Z_GDP, z_inflation and z_unemployment stand for the z-scores of the absolute errors made when reporting the rates of growth, inflation and unemployment, respectively. These z-scores have been calculated by standardizing the absolute errors with respect to the mean and standard deviation of their distribution.

Table A2. Joint regression on the absolute errors (z-scores). Estimated coefficients and standard errors (continued).

Countries	z_GDP	z_inflation	z_unemployment	Countries	z_GDP	z_inflation	z_unemployment
Austria	−0.335***	−0.159***	0.182***	Latvia	−0.086	0.379***	0.070
	(0.051)	(0.048)	(0.047)		(0.061)	(0.058)	(0.057)
Belgium	−0.436***	0.011	1.525***	Lithuania	−0.319***	0.410***	0.206***
	(0.058)	(0.055)	(0.053)		(0.058)	(0.055)	(0.054)
Bulgaria	−0.191***	1.030***	0.336***	Luxembourg	−0.458***	−0.176**	−0.272***
	(0.062)	(0.058)	(0.057)		(0.072)	(0.068)	(0.067)
Croatia	−0.052	0.592***	0.009	Malta	−0.391***	0.056	−0.196***
	(0.055)	(0.052)	(0.051)		(0.071)	(0.067)	(0.066)
Cyprus	0.597***	0.969***	0.204***	Netherlands	−0.705***	−0.138***	0.267***
	(0.077)	(0.072)	(0.071)		(0.052)	(0.049)	(0.048)
Czech Rep.	−0.263***	0.505***	0.906***	Poland	−0.379***	0.419***	0.379***
	(0.054)	(0.051)	(0.050)		(0.052)	(0.049)	(0.048)
Denmark	−0.130**	0.187***	0.201***	Portugal	0.128**	0.577***	−0.128**
	(0.054)	(0.051)	(0.050)		(0.064)	(0.060)	(0.059)
Estonia	−0.443***	0.602***	0.367***	Romania	0.044	0.332***	0.563***
	(0.053)	(0.050)	(0.049)		(0.072)	(0.068)	(0.066)
Finland	−0.272***	0.311***	0.124**	Slovakia	−0.462***	0.140**	−0.357***
	(0.054)	(0.051)	(0.050)		(0.058)	(0.055)	(0.054)
Germany	−0.555***	−0.204***	−0.058	Slovenia	−0.618***	−0.202***	0.146***
	(0.050)	(0.047)	(0.046)		(0.056)	(0.053)	(0.052)
Greece	−0.191***	0.949***	−0.386***	Spain	−0.628***	0.016	−0.164***
	(0.058)	(0.054)	(0.053)		(0.057)	(0.054)	(0.053)
Hungary	−0.404***	0.615***	0.612***	Sweden	−0.454***	−0.119**	−0.185***
	(0.054)	(0.051)	(0.050)		(0.052)	(0.049)	(0.048)
Ireland	−0.347***	0.348***	−0.027	United Kingdom	−0.398***	−0.078	0.423***
	(0.053)	(0.050)	(0.049)		(0.052)	(0.049)	(0.048)
Italy	0.167***	0.479***	0.497***				
	(0.055)	(0.052)	(0.051)				

Breusch-Pagan test of independence: chi2(3) = 4804.98, Pr = 0.0000

Note: See notes under Table 5.

8. References

Akerlof, G.A. 1970. "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84: 488–500. Available at: http://www.jstor.org/stable/1879431.

Akerlof, G.A., W. Dickens, and G. Perry. 2000. "Near-Rational Wage and Price Setting and the Optimal Rates of Inflation and Unemployment." *Brookings Papers on Economic Activity* 1: 1–58. Doi: http://dx.doi.org/10.1353/eca.2000.0001.

Bacchetta, P. and E. van Wincoop. 2005. "Rational Inattention: A Solution to the Forward Discount Puzzle." *NBER Working Paper* 11633. Doi: http://dx.doi.org/10.3386/w11633.

Blendon, R.J., J.M. Benson, M. Brodie, R. Morin, D.E. Altman, D. Gitterman, M. Brossard, and M. James. 1997. "Bridging the Gap between the Public's and Economists' Views of the Economy." *Journal of Economic Perspectives* 11: 105–118. Doi: http://dx.doi.org/10.1257/jep. 11.3.105.

Blinder, A.S. and A.B. Krueger. 2004. "What does the Public Know about Economic Policy, and How Does It Know It?" *Brookings Papers on Economic Activity* 1: 327–387. Doi: http://dx.doi.org/10.3386/w10787.

Bruine de Bruin, W., W. Vanderklaauw, J.S. Downs, B. Fischhoff, G. Topa, and O. Armantier. 2010. "Expectations of Inflation: The Role of Demographic Variables, Expectation Formation, and Financial Literacy." *Journal of Consumer Affairs* 44: 381–402. Doi: http://dx.doi.org/10.1111/j.1745-6606.2010.01174.x.

Bryan, M. and G. Venkatu. 2001. "The Curiously Different Inflation Perspectives of Men and Women." *Federal Reserve Bank of Cleveland Economic Commentary Series* November.

Burke, M.A. and M. Manz. 2014. "Economic Literacy and Inflation Expectations: Evidence from a Laboratory Experiment." *Journal of Money, Credit and Banking* 46: 1421–1456. Doi: http://dx.doi.org/10.1111/jmcb.12144.

Cappellari, L. and S.P. Jenkins. 2003. "Multivariate Probit Regression Using Simulated Maximum Likelihood?" *Stata Journal* 3: 278–294. Stable URL: http://www.stata-journal.com/sjpdf.html?articlenum=st0045.

Cardoso, A.R., A. Loviglio, and L. Piemontese. 2016. "Misperceptions of Unemployment and Individual Labor Market Outcomes." *IZA Journal of Labor Policy* 5: 1–22. Doi: http://dx.doi.org/10.1186/s40173-016-0069-6.

Carrillo, P.E. and M.S. Emran. 2012. "Public Information and Inflation Expectations: Microeconometric Evidence from a Natural Experiment." *Review of Economics and Statistics* 94: 860–877. Doi: http://dx.doi.org/10.1162/REST_a_00213.

Carroll, C. 2003. "Macroeconomic Expectations of Households and Professional Forecasters." *Quarterly Journal of Economics* 118: 269–298. Doi: http://dx.doi.org/10.1162/00335530360535207.

Cavallo, A., G. Cruces, and R. Perez-Truglia. 2016. *Inflation Expectations, Learning and Supermarket Prices. Evidence from Survey Experiments*. Available at: http://www.mit.edu/~afc/papers/Cavallo-Inflation-Expectations.pdf (accessed April 2017).

Curtin, R. 2008. "What U.S. Consumers Know About Economic Conditions." In *Statistics, Knowledge and Policy 2007: Measuring and Fostering the Progress of Societies*, edited by OECD, 153–176. Paris: OECD.

Curtin, R. 2009. "What U.S. Consumers Know About the Economy: The Impact of Economic Crisis on Knowledge?" In Proceedings of the 3rd OECD World Forum on Statistics, Knowledge and Policy: Charting Progress, Building Visions, Improving Life: OECD, 27–30 October, 2009. Busan, Korea. Available at: http://www.oecd.org/site/progresskorea/44129683.pdf (accessed April 2017).

Duffy, D. and P.D. Lunn. 2009. "The Misperception of Inflation by Irish Consumers." *The Economic and Social Review* 40: 139–163. URL: http://hdl.handle.net/2262/58798.

European Commission. 2007. *Special Eurobarometer 67.2. Europeans' Knowledge of Economic Indicators*. Luxembourg: European Commission.

European Commission. 2015a. *Eurobarometer 83.3*. Brussels: TNS Opinion (producer). Cologne: Gesis Data Archive. Doi: http://dx.doi.org/10.4232/1.12356.

European Commission. 2015b. *Europeans and Economic Statistics*. Luxembourg: European Commission.

European Commission. 2015c. *European Economic Forecast. Spring 2015*. Luxembourg: European Commission.

European Commission. 2015d. *Media Use in the European Union. Standard Eurobarometer 82*. Luxembourg: European Commission.

Eurostat. 2016a. *About Eurostat. Overview*. Available at: http://ec.europa.eu/eurostat/about/overview (accessed April 2016).

Eurostat. 2016b. *Glossary: Gross Domestic Product*. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Gross_domestic_product_(GDP) (accessed December 2016).

Eurostat. 2016c. *Glossary: Inflation*. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Inflation (accessed December 2016).

Eurostat. 2016d. *Glossary: Unemployment*. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Unemployment (accessed December 2016).

Eurostat. 2016e. *Information Society Statistics. Internet Activities*. Available at: http://ec.europa.eu/eurostat/web/information-society/data/database (accessed April 2016).

Fonseca, R., K.J. Mullen, G. Zamarro, and J. Zissimopoulos. 2012. "What Explains the Gender Gap in Financial Literacy? The Role of Household Decision Making." *Journal of Consumer Affairs* 46: 90–106. Doi: http://dx.doi.org/10.1111/j.1745-6606.2011.01221.x.

Gal, I. 2002. "Adults' Statistical Literacy: Meanings, Components, Responsibilities." *International Statistical Review* 70: 1–51. Doi: http://dx.doi.org/10.1111/j.1751-5823.2002.tb00336.x.

Giovannini, E., J. Oliveira Martins, and M. Gamba. 2008. "Statistics, Knowledge and Governance." In Proceedings of the Workshop Committing Science to Global Development: Tropical Research Institute, September 29–30, 2008. Lisbon, Portugal. Available at: http://docentes.fe.unl.pt/~lpereira/papers.html (accessed April 2017).

Giovannini, E., M. Malgarini, and R. Sonego. 2015. "What Do Italian Consumers Know About Economic Data? An Analysis Based on the ISTAT Consumers Survey." *Rivista Di Statistica Ufficiale* 3: 25–47. Available at: http://www.istat.it/it/files/2016/06/What-do-Italian-consumers-know-about-Economic-Data.pdf (accessed December 2016).

Gore, S.M., T. Holt, and I.P. Fellegi. 1991. "Maintaining Public Confidence in Official Statistics." *Journal of the Royal Statistical Society. Series A* 154: 1–6. Stable URL: http://www.jstor.org/stable/2982687.

Greene, W.H. 2011. *Econometric Analysis* (7th edition). New York: Pearson.

Holt, T. 2008. "Official Statistics, Public Policy and Public Trust." *Journal of the Royal Statistical Society Series A* 171: 323–346. Doi: http://dx.doi.org/10.1111/j.1467-985X.2007.00523.x.

Jonung, L. 1981. "Perceived and Expected Rates of Inflation in Sweden." *American Economic Review* 71: 961–968. Stable URL: http://www.jstor.org/stable/1803477.

Jonung, L. and D.E. Laidler. 1988. "Are Perceptions of Inflation Rational? Some Evidence for Sweden." *American Economic Review* 78: 1080–1087. Stable URL: http://www.jstor.org/stable/1807167.

Kahneman, D. and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47: 263–291. Doi: http://dx.doi.org/10.2307/1914185.

Klapper, L., A. Lusardi, and P. van Oudheusden. 2015. *Financial Literacy around the World*. Available at: http://gflec.org/wp-content/uploads/2015/11/Finlit_paper_16_F2_singles.pdf (accessed January 2017).

Kunovich, R.M. 2012. "Perceived Unemployment. The Sources and Consequences of Misperception." *International Journal of Sociology* 42: 100–123. Doi: http://dx.doi.org/10.2753/IJS0020-7659420405.

Lucas, R.E. Jr. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4: 103–124. Doi: http://dx.doi.org/10.1016/0022-0531(7290142-1).

Lusardi, A. and O. Mitchell. 2008. "Planning and Financial Literacy: How Do Women Fare?" *American Economic Review* 98: 413–417. Doi: http://dx.doi.org/10.1257/aer.98.2.413.

Malgarini, M. 2009. "Quantitative Inflation Perceptions and Expectations of Italian Consumers." *Giornale degli Economisti e Annali di Economia* 68: 53–80. Stable URL: http://www.jstor.org/stable/41954986.

Mankiw, N.G. and R. Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *NBER Working Paper 8290*. Doi: http://dx.doi.org/10.3386/w8290.

Natcen. 2015. *Public Confidence in Official Statistics*. Available at: http://www.natcen.ac.uk/media/833802/public-confidence-in-official-statistics_-final.pdf (accessed April 2016).

Northern Ireland Statistics and Research Agency. 2015. *Public Awareness of and Confidence in Official Statistics in Northern Ireland 2014*. Available at: http://www.nisra.gov.uk/aboutus/index.html (accessed April 2016).

OECD. 2005. *Statistics, Knowledge and Policy. Key Indicators to Inform Decision Making*. Paris: OECD.

Ottaviani, M.G. 2002. "Statistics, from a Tool for State and Society to a Tool for All Citizens." *International Statistical Review* 70: 30–32. Doi: http://dx.doi.org/10.1111/j.1751-5823.2002.tb00338.x.

Papacostas, S. 2008. "Special Eurobarometer: European Knowledge on Economical Indicators." In *Statistics, Knowledge and Policy 2007: Measuring and Fostering the Progress of Societies*, edited by OECD, 177–196. Paris: OECD.

Reis, R. 2006. "Inattentive Consumers." *Journal of Monetary Economics* 53: 1976–1800. Doi: http://dx.doi.org/10.1016/j.jmoneco.2006.03.001.

Rose, N. 1991. "Governing by Numbers: Figuring out Democracy." *Accounting, Organizations and Society* 16: 673–692. Doi: http://dx.doi.org/10.1016/0361-3682(91)90019-B.

Sims, C.A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50: 665–690. Doi: http://dx.doi.org/10.1016/S0304-3932(03)00029-1.

Souleles, N.S. 2001. "Consumer Sentiment: Its Rationality and Usefulness in Forecasting Expenditures—Evidence from the Michigan Micro Data." *NBER Working Paper* 8410. Doi: http://dx.doi.org/10.3386/w8410.

Stigler, G.J. 1962. "Information in the Labor Market." *Journal of Political Economy* 70: 94–105. Stable URL: https://www.jstor.org/stable/1829106.

Tarde, G. 1903. *The Laws of Imitation*. New York: Henry Holt and Company.

Townsend, R. 1983. "Forecasting the Forecasts of Others." *Journal of Political Economy* 91: 546–588. Doi: http://dx.doi.org/10.1086/261166.

United Nations General Assembly. 2010. *Resolution 64/267. World Statistics Day*. Available at: http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/64/267 (accessed April 2016).

Walstad, W.B. 1997. "The Effect of Economic Knowledge on Public Opinion of Economic Issues." *Journal of Economic Education* 28: 195–205. Doi: http://dx.doi.org/10.2307/1183198.

Walstad, W.B. and K. Rebeck. 2002. "Assessing the Economic Knowledge and Economic Opinions of Adults." *The Quarterly Review of Economics and Finance* 42: 921–935. Doi: http://dx.doi.org/10.1016/S1062-9769(01)00120-X.

Wild, C. 2005. "Education is Everybody's Responsibility." *International Statistical Review* 73: 213–214. Doi: http://dx.doi.org/10.1111/j.1751-5823.2005.tb00274.x.

Received August 2016
Revised July 2017
Accepted July 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1087–1089, http://dx.doi.org/10.1515/JOS-2017-0050

Book Review

Nathan Cruze[1]

Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton. *Modern Data Science with R*. Boca Raton, FL: Chapman and Hall/CRC Press, Taylor and Francis Group, 2017. ISBN 978-1-4987-2448-7, 551pp, $99.95.

Modern Data Science with R originated from materials that supported semester-long introductory and intermediate data science courses offered at Amherst, Smith, and Macalester Colleges. As such, this text is targeted to a rather general audience and assumes only a minimal background in introductory statistics and computer programming concepts as its starting point. The text is more conceptual rather than technical, and the numerous color-coded R code excerpts presented in the text are a useful aid for seasoned and first-time R programmers alike. The upper-level undergraduate data science student and the self-directed learner will both benefit from the contents and layout of the book. In the presented material, the authors draw an important distinction between the sometimes conflated terms "data science" and "big data", and an exhaustive treatment of the latter is outside the scope of the book. While those looking for insights into handling truly massive data sets may be better served by other references, an invested reader should expect to be rewarded with a variety of foundational skills needed to handle, analyze, and interact with a wide variety of medium-sized data in R, i.e., data that may fit on one ordinary computer hard drive while being entirely too big to be stored in memory.

In addition to introducing the R programming language, the text is presented around the use of the RStudio front end, making extensive use of the `knitr` and `markdown` packages. The authors made this decision in the name of enhancing workflow and reproducibility, and emphasize these two points as important corner stones in the practice of data science. An additional R package developed by the authors, `mdsr`, installs most of the additional R packages necessary to carry out the analyses and exercises presented within the text and includes some of the smaller example data sets. A range of interesting larger data sets including airline data, Twitter data, and Federal Elections Committee data are drawn from online sources and referred to repeatedly throughout examples and exercises in the text. Several appendices provide additional information on R, RStudio, and `mdsr`.

Modern Data Science with R is clearly organized along three parts. Part I (Chapters 1–6) covers a variety of introductory topics in data science with an emphasis on data visualization and data wrangling, the process of acquiring and transforming a "raw" data format into a form more suitable for some end use. Exposition and demonstration of the

[1] USDA National Agricultural Statistics Service – Research and Development Division Room 6412A–South Building 1400 Independence Avenue, SW, Washington, District of Columbia 20250, U.S.A. Email: nathan.cruze@nass.usda.gov

former revolves around the `ggplot2` plotting paradigm (Wickham 2016). The latter is focused on data cleaning and data wrangling through the `dplyr` package. For the aims of this text, extensive use of these libraries make sense, although they are certainly not the only available options. While other potentially relevant and useful plotting approaches exist in R, e.g., base graphics (Murrell 2011) or lattice graphics (Sarkar 2008), the extensive and consistent use of the `ggplot2` grammar lowers some of the barriers to understanding other data types and visualizations presented in later supplemental chapters, e.g., mapping spatial data through the use of `ggmap` or `leaflet` packages. The use of `dplyr` (as opposed to `data.table`) naturally fits with many relational database backends, and it agrees with the syntax of other packages demonstrated in the text. In order to get the most out of the remainder of the book, a mastery of the graphics and data wrangling concepts presented in Chapters 3, 4 and 5 becomes essential. The thoughtful and refreshing discussion on professional ethics in Chapter 6 is a welcome addition.

Part II (Chapters 7–10) introduces a number of common statistical concepts and modeling approaches: sampling, bootstrapping, linear regression, statistical learning and prediction, unsupervised learning, and simulation. These are presented lucidly, and in a largely non-technical manner. (Note that a more mathematical treatment of regression models is given in Appendix E.) Entire textbooks have been written on each of these subjects, but the content in Part II may serve as useful introductory material or a supplement to other materials for those unfamiliar with these topics. Part II flows as a natural continuation of the aims and thought processes of the earlier chapters: that data may need structure to be analyzed, that visualization is a powerful tool for recognizing patterns and relationships, that the magnitude and association of a response variable with other explanatory variables may be quantified or predicted (regression, supervised learning), and that patterns in data may still emerge even when an outcome is not explicitly measured (unsupervised learning).

The collection of topics in Part III are essentially independent of the material in the first two parts of the book, but they may be of interest to serious students. Covered topics include the production of interactive data graphics (Chapter 11); SQL, databases, and database administration (Chapters 12–13); and spatial data and mapping, text mining, and network science (Chapters 14–16). At just fourteen pages long, the final chapter titled "Epilogue: Towards 'big data'" can only hint at some of the issues, challenges, and techniques associated with harnessing big data.

A major strength of *Modern Data Science with R* is the shear breadth of topics covered in a non-technical manner, and the generous use of code excerpts to help the reader walk through and reproduce important concepts. There is much to recommend this text to the advanced undergraduate data science students and general R users interested in improving their understanding of data wrangling, data visualization, and discovery.

References

Murrell, P. 2011. *R Graphics*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.

Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer-Verlag.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed. New York: Springer-Verlag.

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1091–1092, http://dx.doi.org/10.1515/JOS-2017-0051

Book Review

Morgan Earp[1]

Francesco Bartolucci, Silvia Bacci, and Michela Gnaldi. *Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata*. CRC Press, 2016. ISBN 978-1-4665-68495, 305 pp, $76USD.

Measurement error is an important aspect of total survey error and is discussed at length in the *Journal of Official Statistics*, but there is very little discussion and/or application of psychometric tools such as Classical Test Theory (CTT) or Item Response Theory (IRT) being used to evaluate measurement error of latent traits (traits that are not directly observed, like respondent burden). CTT and IRT are specifically designed to evaluate the validity and reliability of measures dealing with latent traits, and are commonly used to assess the validity and reliability of psychological and educational measures; however this does not appear to be a common practice in the construction of the survey questionnaires discussed in the pages of *JOS*. In fact, while fifty-five articles come up in *JOS* searching "measurement," only twelve come up searching "reliability," nine searching "validity," six searching "latent," four searching "item response theory," and zero articles come up when using the search terms "Classical Test Theory" or "psychometric." While psychometric assessment is commonly taught in education and psychological research methods programs, it is not generally understood outside of those fields. The authors of the book *Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata,* Bartoulucci, Bacci, and Gnaldi provide a rich and easy to follow overview of psychometric theory that can easily be understood and applied by survey methodologists outside of the fields of education and psychology; they discuss the theoretical framework behind these types of models as well as provide a practical guide for assessing questionnaire latent constructs using psychometric evaluation methods.

Many of the surveys used outside of education and psychology also contain latent traits, traits that we cannot directly observe, but can indirectly measure using a series of related questions (i.e., respondent burden). CTT and IRT can be used to compare and contrast the relationship between items and/or factors to assess and compare the reliability of items or latent construct measures as a whole. CTT assumes that the standard error is the same for all scores in a given population, where IRT assumes the standard errors vary. The authors discuss psychometrics from both a CTT and an IRT perspective. The book is designed specifically for graduate psychometric and statistics courses with an emphasis on measurement via questionnaires, but can be used by any survey methodologist and or practitioner interested in evaluating the reliability of latent construct measurement in their survey. The book is not only rich in theory and provides a thorough background, but it also

[1] Bureau of Labor Statistics – Office of Survey Methods Research, PSB Suite 1950, 2 Massachusetts Avenue, NE Washington District of Columbia 20212, U.S.A. Email: earp.morgan@bls.gov

provides an easy user-guide to building and assessing CTT and IRT models using Stata and R.

The book is organized into six chapters. The first three chapters provide an easy to follow and solid introduction to the psychometric evaluation of questionnaires, including both CTT and IRT, and the subsequent chapters provide both the theory and the programming for working with various types of items and constructs using R and Stata. The first chapter provides an overview of measurement as it relates to item development and questionnaires in the social sciences. The second chapter discusses construct reliability and validity using CTT. Chapters three through six discuss the overall assumptions behind IRT and the varying types of IRT models. Chapter three focuses on dichotomously scored items and discusses the three main types of IRT models including Rasch, two-parameter, and three-parameter models. Chapter four discusses the use of IRT to examine categorical responses including binary, polytomous, and ordinal data. Chapter five compares the different IRT estimation methods as well as goodness-of-fit measures. Lastly, Chapter six discusses other extensions of the IRT model including multi-level/longitudinal data, multidimensional constructs, and ends by comparing IRT to the use of structural equation modeling.

While intended for the psychometrician developing psychological or educational measures, survey practitioners will benefit from using the methods laid out in this text to evaluate the potential for measurement error within surveys and questionnaires designed to measure any latent constructs. CTT and IRT can be used to determine which items have large amounts of measurement error and therefore may not be reliable, and IRT can be used to determine for which persons/households/establishments items may not be reliable, or in some cases invalid given how easy or difficult they are to answer or agree with. Both CTT and IRT can be used to determine if the measurement error of latent constructs is higher for certain groups/types than others using invariance/differential item functioning testing. IRT can also be used to assess the utility of an item scale, to determine which response options distinguish between trait levels (as well as those that do not). Both CTT and IRT can serve as a nice compliment when evaluating the quality of survey items.

This book provides an easy to follow overview of psychometric evaluation methods that can be used to assess questionnaires and surveys designed to measure latent constructs, as well as easy to follow documentation in both R and Stata for the various types of models. This book is of value to both the novice and the experienced psychometrician as well as those who are completely new to the field. With plenty of introduction as well as detail for those who wish to delve deeper into the theory, math, and programming behind the different types of psychometric models, the text serves as an excellent manual for those looking to evaluate latent constructs within questionnaires (whether in the pre-testing phase or generally) or to better understand and review psychometric evaluation research. It provides easy to follow syntaxes and annotated output for both Stata and R, making it a simple and yet rich enough for graduate level students and practitioners alike. This book brings together a vast amount of psychometric knowledge as well as software package programming into a single cohesive book, which is something that has been missing for a long time. Most psychometric books focus on specific types of models or specific software packages, but none to date have provided the type of overview and straight forward explanations, not to mention the number of examples (including syntaxes) that this text does.