

# Redeveloping the Norwegian Household Budget Survey

*- Past, Present and Future*

Kristin Egge-Hoveid<sup>1</sup>, Sverre Amdam<sup>2</sup>

<sup>1</sup> *Statistics Norway, Oslo, Norway; keg@ssb.no*

<sup>2</sup> *Statistics Norway, Oslo, Norway; sad@ssb.no*

## **Abstract**

The paper presents quality aspects regarding the past, present and future design of the Norwegian Household Budget Survey.

The past design was based on paper diary reporting, and an initial and closing CAPI/CATI interview. Coding and editing of data was to a large extent manual. This led to high survey costs. Response-rate below 50 percent, under-reporting and manual processing of data indicated a need to modernize the survey.

At present Statistics Norway is developing a new digitalized survey design to be implemented in 2017. The design is an initial telephone interview, a web diary with automatically coding of COICOP, and a web questionnaire. Paper receipts will be scanned and automatically coded. Digitalized instruments for data collection and more standardized systems for editing will improve data quality in terms of more effective data collection, more precise coding and better editing. More digitalized data collection may, however, influence response rate and lead to more missing or inadequate completion. This places great demands on the technical solutions to be developed. Usability tests and experiments will be conducted in summer 2016.

The future design will explore the use of alternative data sources, such as transaction data from banks, retailers and other financial institutions. These sources may be used both in addition to, and partly instead of a survey. The utilization of such data can reduce response burden and costs, and meet the quality challenge of unit- and item- non-response in a new way. It will also set new standards for the sampling, and volume of data. However, it opens up new challenges such as privacy concerns, technical issues, and dependency on collaboration with big market actors such as banks and retailers.

**Keywords:** Household Budget Survey, Web-Survey, Big-Data

## **1. Background**

Statistics Norway (SSB) conducted its first Household Budget Survey in 1958. In the period 1973 to 2009 data were collected and published annually, and after a three year stay, the survey was last conducted in 2012. The registration period has been two weeks, with data collection spread over one calendar year. The statistics are published in accordance with the standard classification of individual consumption (COICOP).

## **2. The Past – paper diaries and manual coding**

### *2.1 Past data collection design*

The last Norwegian HBS was conducted in 2012, with a gross sample of 7000 households, stratified by region and type of household. A response rate of 48,6 gave a net sample of 3 363 households. The design was threefold; first an initial interview either by home visit or by telephone. The households then registered their expenditures in paper diaries during the assigned two-week period, followed by a closing interview. The closing interview mapped the expenditure on larger and more rarely purchased goods and services, for example purchase of private vehicles, costly articles of clothing, travel expenses and the like. Participants were also asked to collect receipts of all purchases to ensure more complete coverage of expenses.

The HBS 2012 was meant to be primarily a visitor survey, but financial resources limited the number of visits. Over half of all initial interviews and over 40 percent of closing interviews were done by telephone. In retrospect analysis showed that when the initial interviewing were done by telephone, 88 percent completed the closing interview and 83 percent kept a diary, while 95 percent completed the closing interview and 93 percent kept a diary when the initial interview was conducted through visits. This clearly shows the value of personal interviews, but costs are significantly higher.

Because of the overall high response burden associated with HBS there is reason to believe the survey has been encumbered by a relatively high degree of item-non-response. This

might have amplified by the fact that many respondents experienced handwritten diaries as time-consuming and thus an inefficient registration tool.

## *2.2 Processing and editing of data*

The previous Norwegian Household Budget Surveys has relied on extensively manual coding of handwritten diaries, and manual coding of receipts. The managing and processing of data has been automated only to a limited extent, and have therefore been dependent on manual procedures and hard coding of data. The coding procedure and manual handling of data led to inefficient work processes and thus also considerable costs. Manual coding, processing and editing obviously also represent a challenge to the accuracy and quality of data compared to more automated processing routines.

## **3 The Present – digitalized data collection, automatic coding, processing and editing**

In light of the above conditions Statistics Norway is at present developing a new multi-mode and digitalized data collection design to be implemented in 2017. The new design combines an initial telephone interview, a web diary with automatically coding of COICOP, and a web questionnaire to map the larger and more rarely purchased goods and services. Paper receipts will be collected, scanned and automatically coded.

### *3.1 The web diary*

One of the main objectives of the new design has been to give participants the opportunity to fill in their purchases and expenses in a user-friendly online web-diary that can be used regardless of platform; from PCs, tablets and smartphones.

Reduction of costs through more efficient coding has, however, been an equally important objective. Developing a system that automatically categorizes and codes the registered expenditures into the correct COICOP is therefore another important aim of the new HBS. The web diary is therefore developed to integrate with an administrative register of COICOP categories with predefined keywords. This allows the filling of purchases and expenses to be automatically categorized and coded. The system is designed to be self-learning. If a

keyword is not found in the system, coding to the correct COICOP will be done once and will then be stored for all future filings.

Another purpose of the administrative register of COICOP categories is to provide an automatic link between OCR scanned receipts that participants provide from supermarket chains, and Statistics Norway's retail trade registry. For CPI purposes, Statistics Norway collects data from retail trade of food and beverages weekly from the main food chains. The biggest chains cover a significant share of the sales of food and drink in Norway, about 80 percent. To ensure correct mapping of the scanned receipts, we have imported the chains receipt text (bong text) and linked this to the Global Trade Item Number (GTIN). GTIN is in turn coded into COICOP in the retail trade register. All arrears of unknown text or goods will be stored in a separate register, where they can be coded manually.

### *3.2 The Web questionnaire*

In the new design the former CAPI/CATI closing interview will be completed through a web questionnaire integrated with the web-diary. The questionnaire will, as previously, seek to identify more costly and rarely bought goods and services. Integrating the questionnaire with the web-diary permits a flexible completion during the two week period of participation. Earlier the CAPI/CATI interview took about an hour to complete. Making it possible to complete in a flexible way through the recording period might help to reduce response burden and possibly increase the number of completions.

### *3.3 Automated editing and processing of data*

A third important objective of the redeveloping of the HBS has been to improve data processing, editing and production of statistics. These procedures have suffered from much manual handling of data and little flexible processing tools. As a result work processes have been inefficient, costs have been high and data quality has to some extent relied on subjective assessments more than rule-based procedures. The new processing system is not yet fully developed, but the main features will be editing and estimation through the application ISEE (Integrated System for Editing and Estimation) and processing and production of statistics through SAS EG.

### *3.4 Recruitment*

There are many obvious reasons to redesign the HBS survey to be more cost efficient and to utilize more modern and efficient tools for collecting data. Statistics Norway's strategy explicitly states that we shall develop and adopt more digitized collection methods.

With the new design we are quite confident that the data collection will be more efficient and that we will get better data quality through better and more accurate coding and more automated procedures for editing, estimation and production of statistics. However, we do not yet know whether we will manage to achieve the goal of minimum 40 percent response, with the new design. There are several challenges associated with the transition from personal interviews and follow-up to a fully web-based filling. Other statistical agencies, including the Netherlands, have experienced a lower response rate in the transition to web-based filling.

An important prerequisite for success with the web-based solution is that it is highly user friendly, and that technical errors do not occur. We have conducted several user tests underway and we will carry out a pilot of the survey in June 2016 to test all the instruments. Equally important, however, is to have a solid plan for recruitment and follow up of the respondents. Among other things we have developed a short one minute animated instruction video, easy and comprehensive respondent materials and pop up notices in the diary to remind the respondents to record items easy to forget. We will also use a system for automatic sending of emails and SMS as follow up procedures if not logged in, if filling is incomplete etc. Personal telephone follow up will be used if needed. Finally we have planned with a monetary incentive which can be redeemed when the survey is completed satisfactorily. We also believe the initial CATI interview/recruiting, with highly skilled and experienced interviewers, is important to recruit and keep respondents.

## **4. The Future - Using digital transaction data in HBS statistics**

As a part of the HBS-renewal project, Statistics Norway has done an initial investigation of how to use digital transaction data in household budget statistics. More specifically we

looked into which sources that exists, which institutions possess these data, and further how one might utilize these data in the future. This work gave insight, but also raised some big challenges.

In Norway 94 per cent of all household transactions and purchases are done electronically (payment by card or by internet-bank) (Norges Bank 2014). Each such transaction is registered either by a payment processor and/or by a bank. In addition retailers and chains increasingly store detailed information about customer's purchases through loyalty programs, and other customer attachments. Gaining access to such data sources, and finding ways of processing them for use in HBS statistics, can add, and in time even replace parts of the present survey design, reducing both costs and improving quality in HBS statistics.

#### *4.1 Data sources*

When looking at digital transaction data relevant for household budget statistics, it is meaningful to speak of two principal data sources: "Bank transaction data" and "Bong data". In addition one is dependent on the statistical office's basis population registers to provide frames of person and household samples.

Each electronic bank transaction to/from an account is registered by the respective bank and/or payment processor. These data contain information about the total sum of the transaction, transaction date, bank account number, and addressee (who receives/send the sum). The advantage of these data is that the linkage to individuals is established through the account and/or card number. Also, these data cover all digital transactions, and is (in Norway) stored by a handful of payment processors that service all Norwegian banks. The disadvantage of these data is the lack of details in each transaction. One cannot break the data into specific goods/services, at least not in accordance to the COICOP standard. Bong data is stored in the sales terminal system of each retailer and contain detailed information about the purchase. Bong data can be described as receipt data, containing information on price, quantum and product names of each bought product. The upside is that the level of details within these data meets the COICOP-standard and more. The downside is that only a handful of retailers have linkage to individuals, as the card/account number is not stored in

the sales terminal systems by default. Some retail chains in Norway have a linkage through customer loyalty programs, where members register their payment cards, but most don't.

To sum up you have one source of data which has good coverage of household transactions but lacks the details required, and one source that has sufficient details but lacks in coverage. In spite of the limitations in each of these sources, there are several ways in which they both can supplement and/or replace more traditional data collection methods.

#### *4.2 Using digital transaction data for HBS purposes*

Grocery expenditures represent the most burdensome part of the HBS today, both in terms of volume and detail level, thus this is an area in where the cost/benefit for alternative sources would have good potential. In Norway two of the largest grocery retailer chains (GRCs), covering approximately 75 % of the grocery market, offer loyalty programs which record the members purchases. When accumulated these files form a rich source of data. By linking social characteristics to those present in the retailer's database, the national statistical office (NSO) can produce statistics analogue to that produced from the survey samples. The data can be accumulated over time to avoid seasonal variation associated with sample diary report. The large amount of available data also means that sampling uncertainty reduces error in this source. Another advantage is that the linkage between bong data (same data as scanned receipts) and COICOP classification is established through the 2017 survey. The key quality issues in this kind of data are rather selectivity and potential coverage errors. The loyal GRC customers is unlikely a simple random sample of the target population of the HBS. Not all the household members of the loyal GRC customers may be found in the same database. Retails at GRCs may not cover the whole range of retail expenditure. Methodological issues to be explored are alternative possibilities of detecting and adjusting for such errors. Depending on factors such as these, one further needs to find how these bong data best can be used. One possibility is to use the data as auxiliary data alongside a traditional survey, to reduce errors in the observed sample. Another alternative is to use bong data as an input mode in the HBS survey sample, introducing bong data from GRC members in the target sample to increase data volume and accuracy in the sample data. Thirdly one might replace survey data for grocery consumption in the HBS, using bong data

alone to measure the consumption of groceries among Norwegian households. This will clearly have most impact on response burden and survey costs, but also set the most stringent requirements on the capability of the bong data.

#### *4.3 Bank transaction data*

As the level of details in this source is insufficient in itself, and there is no established link to bong data, one has to depend on some way of specifying, or detailing transactions into respective goods/services. One way of doing this is to introduce the bank transaction data in a web survey design, where respondents can use their transaction log, for example as shown in the internet bank, as a basis for their web diary completion. This could enrich data volume and enhance quality, but one still would be fully dependent on a survey. Also, one would have to invest in technical integration with banks and similar, and there is reasons to believe that far from all respondents would use the tool due to privacy concerns. The cost/benefit for such a solution may be difficult to justify. Another possibility may lie in so called Private Financial Management (PFM) solutions, where the bank itself specifies many of its customer's transactions into consumer categories as a tool for customers to overlook their own economy. Today the banks PFM systems in Norway are not developed at a detail level sufficient for HBS purposes, but this is an area of continuous development for many large banks. One hybrid solution could be to collaborate with banks to use the PFM solution in the internet bank, as a web diary platform, and further let a sample of bank customers specify more accurate on transactions where that is needed. This might reduce development costs, and could also counteract privacy concerns among the respondents. However, this requires close collaboration and willingness within the bank sector.

Gaining access to these sources of data, and find ways to integrate them in HBS statistics in a cost efficient way needs to be further investigated. There are several methodological challenges, as well as technical challenges and privacy concerns. To meet the collaboration between NSOs and private institutions such as banks, payment processors and GRCs seems of great importance.



## **5. References**

Norges Bank (2014): Utviklingstrekk i kunderetta betalingsformidling - 2013 (Trends in customer-oriented payments). Oslo, Norges Bank Memo no 1 2014.