Statistical Journal of the IAOS 31 (2015) 297–304
DOI 10.3233/SJI-140863
IOS Press

A strategy to test questionnaires at a national statistical office

Andreas Persson[a,*], Anette Björnram[a], Eva Elvers[b] and Johan Erikson[a]
[a]*Statistics Sweden, Örebro, Sweden*
[b]*Statistics Sweden, Stockholm, Sweden*

Abstract. There are many methods (for example, expert reviews, cognitive interviews, and experiments) to test questionnaires and other data collection instruments. In practice, all surveys cannot be tested with all methods; there has to be a balance with regard to survey importance, consequences of data errors, available resources, and costs. Statistics Sweden has developed a strategy how to test questionnaires in different surveys, taking the abovementioned factors into account. This strategy is based on a small set of survey characteristics which mainly are taken from a database with classifications for many surveys. Based on how a survey is classified in the chosen characteristics, the strategy proposes different levels of testing. These levels vary in both ambition and the methods included. The strategy has, since implemented, increased the amount of testing but in a differentiated way, based on risk and resources.

Keywords: Questionnaire testing, pre-testing, risk management, survey planning, data quality

1. Methods to test questionnaires

The questionnaire and other, similar, data collection instruments play a fundamental role in the production of survey statistics. The questionnaire is the tool via which the data is collected from the respondents. Flaws in the questionnaire can have negative consequences in many different areas. Hence, it is important to have a well-designed questionnaire. However, writing questions is difficult. There are several books on question design [7,15] but the knowledge on how to consistently write good questions is still far from complete. Testing and evaluation of the questionnaire is therefore important to identify problems with questions, improve the questionnaire and, in that way, reduce measurement error. In addition, evaluating the questionnaire can improve the administration of the questionnaire, making it easier and smoother to complete. In interview surveys, administration time is related to costs and interruptions due to a flawed questionnaire might also influence the measurement negatively [2]. In self-administered questionnaires, improved administration should mean reduced respondent burden. This is important to avoid fatigue and undesirable response styles such as satisficing [9] (which both are linked to data of lower quality) and to promote participation in future surveys [16]. Thus, it is important to test the questionnaire in advance.

An important question is, then, *how* should the questionnaire be evaluated? Survey methodology and practice have established a number of different methods for this purpose [10,12]. Some examples are expert reviews, cognitive interviews with probing, vignettes and think-aloud protocols [17], debriefings with interviewers or data editors, monitoring and behaviour coding of interviews [6], and analysing quantitative data, with or without experimental design, to evaluate individual questions or the questionnaire as a whole [1]. These methods differ in many ways and how they best should be applied or combined is still a rather open question [3,4,11,18,19]. However, given that the findings from different methods only seem to correlate to a degree [4,18,19] using more test methods should, in general, detect more and a greater scope of potential prob-

*Corresponding author: Andreas Persson, Statistics Sweden, 701 89 Örebro, Sweden. Tel.: +46 19176248; E-mail: andreas.persson@scb.se.

lems with the questionnaire than fewer methods would. As such, evaluations of a questionnaire should benefit from applying more than one method. In line with this, ambitious re-designs of questionnaires in surveys often use more than one evaluation method [8,14].

2. Questionnaire testing at a statistical agency – limited budgets and repeated surveys

In practice, at a statistical agency further factors than methodological ones usually play a role in the choice of evaluation methods. Time and financial resources often put restraints on the evaluation of the questionnaire. Surveys have fixed budgets and even though questionnaire evaluation is important, it is still only one of many important factors in good survey design, all competing within the same survey budget. In practice at a statistical agency, one has to consider costs, benefits, and the big picture – the total design of the survey. As such, evaluating the questionnaire is important (see above) but perhaps not always or indiscriminately. The total design, the benefits and costs of questionnaire testing, and the budget of the survey have to be considered.

Moreover, different surveys have questionnaires that differ in status. One important difference, among many, is whether the questionnaire is new (i.e. draft status) or established and possibly already evaluated. Questionnaire evaluation obviously does not fill the same purpose in these situations. Such factors should influence in what way, to what extent, and perhaps even whether the questionnaire should be evaluated at all, given the perspective of total design and set budgets. For instance, a statistical agency usually has many repeated surveys that are conducted on, for example, monthly or annual basis. The demands on the questionnaires in such surveys should be high since they are used repeatedly. The continuous data collection in repeated surveys also permits questionnaire evaluation in a more cyclic, long-term fashion – where information from previous rounds is used for improvements to future rounds. However, from a practical perspective, it is not rational to continue to evaluate an already evaluated questionnaire unless there are indications (for example, from the production process) that suggest that it is needed. This, to balance needs and resources, seems to be a common challenge at many statistical offices [13]. Thus, different questionnaires, for example those in new and repeated surveys, merit different approaches concerning testing.

3. Situations when testing is needed

The European Statistics Code of Practice states that "In the case of statistical surveys, questionnaires are systematically tested prior to the data collection" [5]. In practice, the conditions that merit evaluations are not given. Recently, Statistics Sweden has developed a set of conditions that stipulates whether a survey's questionnaire should be evaluated before data collection or not. The main four situations where testing is required are:

- The questionnaire is new.
- The questionnaire has been changed, for example by adding new questions or a new data collection method.
- The context has changed in ways that could influence the measurement (for example, if the questionnaire should be used for a different population).
- There are indications of problems with the questionnaire (based on process data, logs, debriefings, or other sources).

Thus, if any of the above conditions are valid, the questionnaire should be tested before the data collection.

4. The extent of testing

Conditions such as those mentioned determine whether the questionnaire should be evaluated or not. However, given that a questionnaire should be tested, how extensive should the testing be? As shown above, the answer should be different for different surveys, depending on, among other things, the status of the questionnaire and characteristics of the survey. Moreover, resources and total design have to be considered. As such, how extensive the testing should be appear to require some individual investigation for each survey, based on the specific conditions at hand.

Unfortunately, such survey-specific assessments do not correspond very well to the large-scale production of statistics at a statistical agency where hundreds of surveys are conducted every year in a steady stream, where in-house communication can be challenging, where there is a conflict of resources, where the measurement error has not always been of highest priority, and where both the survey manager and the cognitive lab have to plan and allocate resources for testing well in advance. In contrast, ideally there would be an explicit strategy which proposes different evaluations for

different surveys and, in that way, facilitates the testing process.

To overcome the problems outlined above, it seems that such a strategy cannot be survey-specific but must operate on a more general level. That is, the strategy must discriminate between different surveys' needs concerning resources and methodological issues but not to the extent that it becomes too complicated or too complex to communicate and apply in the regular production. Such a strategy should help both the survey managers and the cognitive lab in planning for questionnaire testing. Although such a general strategy undeniably would mean standardisation, with the accompanying disadvantages of not acknowledging uniqueness, it should just as well promote that questionnaire testing becomes a part of each survey's plans and not a last-minute resort. Thus, an explicit strategy for questionnaire testing should facilitate questionnaire testing at a statistical agency and in turn improve the measurement.

5. A test strategy developed at Statistics Sweden

To promote that questionnaire testing becomes a part of the plans of the surveys, Statistics Sweden has developed such a general test strategy. One goal was that the strategy should be intuitive and easy to communicate and apply in the production. Another goal was that the strategy should, without being too complex, take survey characteristics into account and assign different surveys to different levels of testing. Concerning resources, the strategy has to be rational and not, for example, propose major testing for a survey of minor importance. Another question is then how to determine survey importance.

5.1. A risk-based reasoning

How should surveys be differentiated? One perspective is that of risk. A risk consists of the factors likelihood and consequences. The likelihood of flaws in the questionnaire is difficult to estimate in advance (especially with new questionnaires). The consequences, however, can be better forecasted. Flaws in a questionnaire can have a negative impact on the respondents, the data collection and editing and, in the end, the quality of the statistical output. The consequences of flaws in the questionnaire thus depend on the impact of the statistics from the survey. If the statistics are used widely and as a basis for important decisions, the consequences could be severe, in comparison to the opposite conditions. Hence, flaws in the questionnaire have different consequences for different surveys. Great consequences should therefore merit more extensive testing to reduce the likelihood of problems occurring in the questionnaire. Hence, a test strategy should discriminate whether flaws in the questionnaire are likely to have minor or major consequences. How, then, should this reasoning be applied in a test strategy?

5.2. Characteristics of surveys

When working on the strategy and its implementation at Statistics Sweden, three broad types of surveys with questionnaires were relevant, expressed below in terms of the type of statistics produced:

1. Regular, appropriation financed statistics, for which Statistics Sweden is responsible.
2. Fee-financed statistics produced by Statistics Sweden: mostly regular, appropriation financed statistics, for which another government agency is responsible.
3. Fee-financed statistics produced fully or partly by Statistics Sweden: often one-time surveys.

The first type (1) was a natural starting point since Statistics Sweden fully control these surveys whereas the paying customer has an influence on the other types.

We used Statistics Sweden's database of surveys for information to discriminate between different surveys. The database has many variables which are used in systems for publishing statistics, metadata, and economic administration. It covers official statistics from Statistics Sweden and all other responsible agencies and other regular statistics from Statistics Sweden. Several characteristics were studied. A few were omitted due to being irrelevant or strongly correlated with those chosen. In the end, we chose three characteristics which all capture how severe consequences flaws in the questionnaire might have. Together they give twelve (2*2*3) possible categories for surveys. The characteristics are:

- Official statistics (yes or no):
  The Official Statistics Act states that official statistics are statistics for public information, planning and research purposes in specified areas produced by appointed government agencies in accordance with the provisions issued by the Government. Official statistics shall be objective and made available, free of charge, to the public.

- Importance for society (yes or no):
  The survey has this characteristic if its content is considered important in order to avoid or handle a critical situation for society or during times of alert.
- Importance of correctness (three categories):
  This characteristic includes what harm incorrect information can cause in terms of, for example, errors in decisions, reduced confidence, and costs due to breaches of contract (and in general). The categories are (1) moderate harm, (2) considerable harm, and (3) serious harm due to the incorrectness.

Thus, these three characteristics are all relevant when considering risks and, consequently, the amount of testing that different surveys merit. For example, a survey which statistics are official statistics, important to society and have a high importance of correctness should be tested more extensively, since the consequences of flaws in the questionnaire can be severe. Such surveys should therefore be assigned to high levels of testing. With the same logic, surveys with the opposite classifications should be assigned to low levels of testing.

5.3. Levels of testing

For the first type of surveys – for regular, appropriation financed statistics for which Statistics Sweden is responsible – the test strategy uses the three characteristics just described. We investigated how many and which surveys that were assigned to different combinations of the three characteristics. The aim was to get a balanced distribution of surveys considering risks, testing, and resources. We found that three testing levels (B, C, D in increasing order) matched these needs. Hence, each of the twelve combinations (2*2*3) were assigned to level B, C, or D. This assignment resulted in about 50% of the surveys on the lowest testing level (B), 40% on the middle testing level (C), and 10% on the highest testing level (D). This means that the highest testing level is used for about ten surveys.

Hence, how a survey is classified in the survey characteristics determines a level of testing for that survey (B-D). It should be noted that this assigned level represents a minimum in our test strategy. The survey manager and the management team of the survey can decide to test on a higher level.

Many of the surveys of the second type – regular surveys for appropriation financed statistics for which another government agency is responsible – are included in the same database, and therefore the same characteristics (Official statistics, Importance for society, and Importance of correctness) are used to determine their level of testing as well. In addition, we added a further, lower testing level (A), which is not used for the surveys where Statistics Sweden is responsible for the statistics. This level is for surveys with the lowest value on each of the three characteristics. In practice, most surveys are assigned to higher levels than A. Statistics Sweden recommends surveys to test their questionnaire on their assigned level but in these cases the customer has the final call. They might choose to test on a lower level or to not test at all. The level A is always a minimum, however, and is included in the price.

The surveys of the third type – for other fee-financed statistics produced fully or partly by Statistics Sweden – are often ad hoc in character and not featured in the database. As such, other characteristics are needed to categorise these surveys. Examples of characteristics that are reasonable to consider are the sensitivity of the topic and the size of the survey. We have made a tentative categorisation for these surveys but it has rarely been applied in practice so far.

Statistics Sweden is certified to the international standard ISO 20252 for market, opinion and social research. The test-levels have all been designed so that they fulfil the ISO 20252 standard's requirement on pre-testing.

5.4. Testing combinations

Figure 1 below shows the test strategy with eight testing combinations, one combination per box in the figure. The combinations are based on two aspects. First, there are two different situations, depending on the availability of prior data. These situations with and without prior data to analyse are labelled N and P for new and previous, respectively. They are shown in the upper and lower row of the figure. Second, there are the testing levels, which are the columns in the figure (A-D). Thus, the test strategy consists of eight combinations, each with one or more testing methods, as shown in the boxes.

As explained above, the upper part of the figure shows the situation when there is no prior information available. This is primarily relevant for new surveys or new questionnaires. In such cases, methods that require already collected survey data (for example, analyses of survey- or process data, record checks, recorded interviews or debriefings with interviewers) cannot be applied unless a pilot study is conducted (see combi-

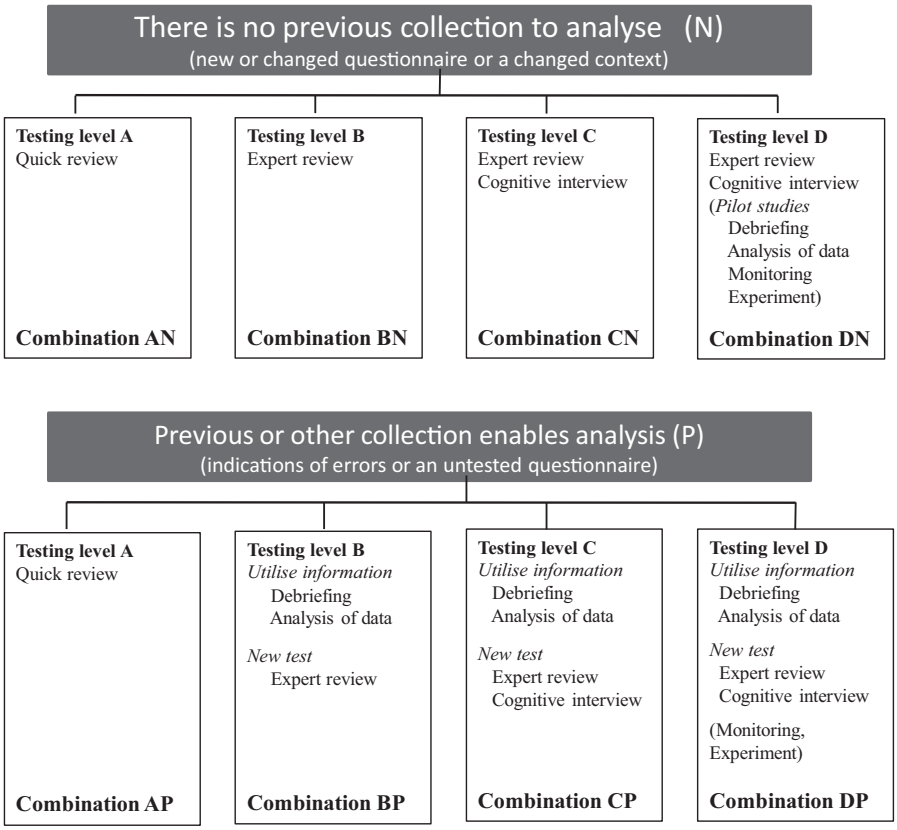

Fig. 1. Testing combinations for different levels (A-D) and depending on whether there is a previous collection to analyse or not.

nation DN). The sequence (from A to D) represents an increase in testing, both concerning the number of methods included and how resource-intensive they are. When there is data available from a previous collection (as in repeated surveys), there are more possible methods, as the lower part of the figure shows. On the highest testing level, D, the survey manager and the management team have to make an appropriate choice from the list in the figure together with the cognitive lab, based on the perspective of survey needs and total design.

Concerning the methods in the figure, a quick review is a screening of potential problems in a questionnaire. It follows a checklist, identifies problematic questions and include solutions to simple, but not to difficult, problems. An expert review is a more extensive review that includes at least two experts on questionnaire design and also includes solutions to most problems. Statistics Sweden's cognitive interviews are based on probing techniques and include 7–10 test interviews. Other agencies may use other methods than these or conduct the same methods in different ways. The methods included are those that fit Statistics Sweden and our in-house competence. The test strategy does not suggest how to optimally combine test methods but how to best distribute finite test resources among surveys that differ in risk. Other agencies might be better off using their test resources in other ways than in the figure, based on their methodological strengths and weaknesses. Given this, the mix of methods in our test strategy follows some general ideas such as to target measurement problems from different perspectives by, where possible, mixing qualitative and quantitative methods, or empirical methods and those based primarily on individual judgments.

6. Implementation of the test strategy

The test strategy was developed at Statistics Sweden over a long time period. A small core group was responsible, but many co-workers served as critical and constructive readers from different points of view. When there was a preliminary version of the test strategy ready, a more formal way was used to communicate within Statistics Sweden. The test strategy was re-

ferred to the most affected departments, i.e. the data collection and the subject matter departments, and also to the group working on the implementation of the ISO 20252 standard. The feedback from the consultation was very supportive and with many constructive comments. This improved the result, and it was also an important step in communicating the strategy and giving it an in-house legitimacy.

A formal decision was taken by the Director General in October 2010. This decision included the test strategy itself and also an implementation plan. Even if some surveys were already tested, at least to some extent, the new strategy proposed increased testing of surveys overall. All surveys could not attain their assigned level immediately. Considering the resources of the cognitive lab and of individual surveys, the strategy had to be implemented successively over time. The sequence and pace of testing were a part of the implementation plan for the following few years, again based on a risk perspective. New surveys and redesigned surveys were to follow the strategy directly, from the beginning of 2011.

When survey managers plan for the next calendar year they should consider the strategy and whether the survey has reached the assigned level or not and, if needed, allocate resources and schedule testing of the questionnaire.

7. The results so far

What are the results of the test strategy so far? It is not given how to best evaluate this since quantitative data is lacking in many cases (for example, concerning the measurement error). The presentation below focuses on experiences, for instance of the question designers and on logs from the cognitive lab.

7.1. The role of the question designer

The strategy has had a positive impact in many ways. First, since the test strategy is mandatory to follow, questionnaire testing is now a given part of the production process. It is something that every survey manager has to acknowledge and act according to. As a consequence, the question designers from the cognitive lab have received a greater authority, and they are now less of a peripheral service and more of a close partner for collaboration. Second, with this expanded role the question designers have been involved in broader work which might not classify as question design or testing per se but in which the question designers' competences in cognitive psychology and communication are very beneficial, such as in the survey's overall communication strategy with the respondents. Overall, it seems as if the question designers have got an expanded role and that this has benefited the surveys.

7.2. The amount of testing

Since one of the goals of the test strategy was to increase the amount of questionnaire testing overall, one way to evaluate the test strategy is to investigate whether the questionnaire testing has increased or not. For this analysis we used information from the cognitive lab's logs which include documentation on questionnaire testing from 2007 and onward.

The quick review is an interesting method for this purpose. According to the test strategy, this test method is only valid on the lowest test level and for fee-financed statistics produced fully or partly by Statistics Sweden for a paying customer. Statistics Sweden encourages higher level of testing but if the customer is not interested in paying for questionnaire testing specifically, such surveys are assigned to level A and a quick review (level A is mandatory, included in the price and cannot be rejected). Therefore, the surveys tested with a quick review would, in general, not have been tested at all previously since it is primarily used for customers that do not want to pay for extra questionnaire testing. There are exceptions, for example customers that would have chosen another pre-testing method if the quick review was not included in the price, but generally this should be true. During 2011 and 2012, the two following years after the test strategy was implemented, the cognitive lab conducted 115 quick reviews. That is, many surveys got feedback on their questionnaire from question designers that they would not have got previously.

Another way of evaluating the test strategy is to look at the regular, appropriation financed statistics, for which Statistics Sweden is responsible. The test strategy assigned those surveys to level B, C, or D, depending on how the survey was classified in the three chosen survey characteristics. In 2011, when the test strategy was introduced, 25 surveys fulfilled their assigned level of testing. Two years later that number had doubled, 50 surveys had reached their assigned level. In addition, in 2011, 39 surveys had not done any questionnaire testing according to the cognitive lab's logs. In 2013 the corresponding number had been reduced to 12 surveys. This suggests that the test strategy has

been successful in facilitating questionnaire testing in general and differentiated questionnaire testing, based on risk, in particular.

Taken together, the test strategy seems to have enhanced questionnaire testing in general, albeit more so for those surveys that were assessed as more important.

8. Conclusions

Statistics Sweden's test strategy has several strengths. Since it is based on given classifications, it makes it possible for survey managers to plan for testing well in advance. In addition, the strategy is rational – limited resources are used where they are best needed, based on risk. Moreover, the classifications are not new for this test strategy but are taken from a database that many co-workers are already familiar with. The strategy therefore includes relatively simple principles, which both the cognitive lab and other staff can grasp and follow. In addition, the many co-workers and departments involved in the developmental work assure that the strategy has taken many perspectives into account and fits the big picture.

Recently, there has been an increased interest in how different test methods perform or compare [3,11,19]. The test strategy does not contribute to that line of research but concerns how to best distribute finite test resources in practice at a large agency. Even though the test strategy is based on general ideas concerning how to mix test methods (to combine qualitative and quantitative data, or empirical methods with those including primarily individual judgment) the proposed combinations of methods might not be optimal for specific surveys or for other agencies. However, the main goal was not to present optimal testing from a methodological view for each survey but to facilitate and establish a baseline for testing in general and to distribute resources rationally (from a risk perspective). Thus, this test strategy suggests how to best distribute test resources among different surveys. How these resources are best spent might differ between different agencies depending on, for example, the agency's methodological orientation or competence.

On the whole the implementation of the strategy has been successful. There are several reasons for this. First, the implementation has been strengthened by the role of the methodologist in the survey management team. Their role is, among other things, to consider whether the questionnaire should be tested or not, as mentioned above, according to a routine in the annual planning and to then reserve the corresponding cognitive-lab resources for the next year. This ensures that questionnaire testing is not overlooked in the survey planning Second, there is, in general, an increased understanding of the importance of well designed questionnaires and of the risks of measurement errors. Third, the test strategy is also well in line with other recently introduced work at Statistics Sweden. For instance, the departments make risk analyses in the annual planning, where different activities are judged based on the risk of problems. Thus, the test strategy was timely and fitted well with other events in the organization. This contributed to the successful implementation.

Statistics Sweden has the advantage of having a database with survey characteristics. This was a good starting point for the development of the strategy. However, the ideas and principles do not depend on these but can be used without pre-made classifications, and, hence, by any statistical agency. Nevertheless, we believe that it is important to choose relatively objective characteristics and to avoid too complex systems when differentiating different surveys.

This strategy has received considerable attention at Statistics Sweden. This is likely due to the strategy proposing that questionnaires are tested in a structured, differentiated, and well motivated way that also makes planning in advance possible. Our evaluation of the strategy suggested that the test strategy has been successful in emphasizing and increasing questionnaire testing overall but also in a differentiated way. Thus, so far, the strategy seems to work well in facilitating questionnaire testing and distributing the resources where they best are needed.

References

[1] P.P. Biemer, *Latent class analysis of survey error*. NJ: Wiley, 2010.

[2] P.P. Biemer and L.E. Lyberg, *Introduction to Survey Quality*. NJ: Wiley, 2003.

[3] F.G. Conrad and J. Blair, Sources of Error in Cognitive Interviews, *Public Opinion Quarterly* **73** (2009), 32–55.

[4] T.J. DeMaio and A. Landreth, Do different cognitive interview techniques produce different results? in: *Methods for Testing and Evaluating Survey Questionnaires*, S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer, eds, Wiley. NJ., 2004, pp. 89–108.

[5] Eurostat, *European Statistics Code of Practice for the National and Community Statistical Authorities – revised edition*. Adopted by the European Statistical System Committee 28th September 2011, 2011.

[6] F.J. Fowler, Coding the behavior of interviewers and respondents to evaluate survey questions, in: *Questions Evaluation Methods*, J. Madans, K. Miller, A. Maitland and G. Willis, eds, Wiley, NJ, 2011, pp. 7–21.
[7] F.J. Fowler, *Improving Survey Questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications, 1995.
[8] D. Giesen and T. Hak, Revising the Structural Business Survey: From a Multi-Method Evaluation to Design, *CBS report*, 2007.
[9] J.A. Krosnick and D.F. Alwin, An evaluation of a cognitive theory of response-order effects in survey measurement, *Public Opinion Quarterly* **51** (1987), 201–219.
[10] J. Madans, K. Miller, A. Maitland and G. Willis, eds, *Question Evaluation Methods*: Contributing to the Science of Data Quality. Wiley; NJ, 2011.
[11] K. Olson, An Examination of Questionnaire Evaluation by Expert reviewers, *Field Methods* **22** (2010), 295-318.
[12] S. Presser, M.P. Couper, J.T., Lessler, E. Martin, J. Martin, J.M. Rothgeb and E. Singer, Methods for testing and evaluating survey questions, *Public Opinion Quarterly* **68** (2004), 109–130.
[13] S. Sattelberger and K. Blanke, Between demand and reality: Ensuring efficiency and quality in pretesting questionnaires. *Proceedings of European Conference on Quality in Official Statistics – Q2012*, 2012.
[14] N.C. Shaeffer and J. Dykema, A Multiple-Method Approach to Improving the Clarity of Closely Related Concepts, in: *Methods for Testing and Evaluating Survey Questionnaires*, S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin and E. singer, eds, Wiley. NJ., 2004, pp. 475–502.
[15] R. Tourangeau, L.J. Rips and K. Rasinski, *The Psychology of Survey Response*. Cambridge University Press, 2000.
[16] M. Wenemark, A. Persson, H. Noorlind-Brage, T. Svensson and M. Kristenson, Applying Motivation Theory to Achieve Increased Response Rates, Respondent Satisfaction and Data Quality, *Journal of Official Statistics* **27** (2011), 393–414.
[17] G.B. Willis, Cognitive Interviewing: A tool for improving questionnaire design. Thousand Oaks, CA: Sage, 2005.
[18] G.B. Willis, S. Schechter and K. Whitaker, A comparison of cognitive interviewing, expert review, and behavior coding. What do they tell us? *Proceedings of the Section on Survey Research methods*, American Statistical Association, 1999, 28–37.
[19] T. Yan, F. Kreuter and R. Tourangeau, Evaluating survey questions: A comparison of methods, *Journal of Official Statistics* **28** (2012), 503–529.