# CONTENTS

**Metodološki zvezki, Vol. 12, No. 1 and 2, 2015**

*Anuška Ferligoj*
How to Improve Statistical Literacy? 1

*J. A. Diaz-Garcia and F. J. Caro-Lopera*
Asymptotic Normality of the Optimal Solution in Multiresponse Surface Mathematical Programming 11

*Gregor Sočan*
Empirical Option Weights for Multiple-Choice Items: Interactions with Item Properties and Testing Design 25

*Nejc Berzelak, Vasja Vehovar and Katja Lozar Manfreda*
Web Mode as Part of Mixed-Mode Surveys of the General Population: An Approach to the Evaluation of Costs and Errors 45

*Denis Marinšek*
A Review of Capital Structure Theory Using a Bibliometric Analysis 69

*Juergen H.P. Hoffmeyer-Zlotnik and Uwe Warner*
Design and Development of the Income Measures in the European Social Surveys 85

*Antonio Angelo Romano, Giuseppe Scandurra and Alfonso Carfora*
Environmental, Generation and Policy Determinants of Feed-in Tariff: a Binary Pooling and Panel Analysis 111

**Metodološki zvezki, Vol. 12, 2015**

## Reviewers for Volume Twelve

Rok Blagus
Andrej Blejec
Gaetano Carmeci
Patrick Doreian
Anuška Ferligoj
Davide Fiaschi
Herwig Friedl
Angelika Geroldinger
Mitja Hafner Fink
Nataša Kejžar
Damjana Kastelec
Irena Krizman
Slavko Kurdija
Giovanni Millo
Irena Ograjenšek
Lorenc Pfajfar
Jože Rovan
Damjam Skulj
Michèle Ernst Stähli
Janez Stare
Gaj Vidmar
Aleš Žiberna

Metodološki zvezki, Vol. 12, No. 1, 2015, 1-10

# How to Improve Statistical Literacy?

Anuška Ferligoj [1]

**Abstract**

In the first part of the paper current initiatives and latest publications with several ideas and good practices for improving statistical literacy are highlighted. In the second part some recommendations for the main actors dealing with statistics are offered. These actors are: educational institutions, statistical offices and other statistical institutions, statistical societies and media. It is pointed out that the cooperation of these actors is essential for improving statistical literacy.

## 1 Introduction

At the beginning of the twentieth century H. G. Wells wrote: "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write". How true! Lancaster (2011) concluded in *How Statistical Literacy, Official Statistics and Self-directed Learning Shaped Social Enquiry in the 19th and Early 20th Centuries* that we are still striving to achieve this in the modern technological age of the twenty-first century. Understanding statistical concepts and methodologies is essential for the proper and efficient use of statistical data collected and published by statistical offices and other authorised institutions. To ensure the better use of statistical data much effort must be put into improving statistical literacy in society.

## 2 What is statistical literacy?

While there are several definitions of statistical literacy, most are based on the definition given by Katherine K. Wallman (1993) in the speech she delivered when she became President of the American Statistical Association: "Statistical literacy is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions".

Gal (2002) introduced two components of adult statistical literacy: knowledge elements and dispositional elements. The former deals with people's ability to interpret and critically evaluate statistical information, data-related arguments or stochastic phenomena they may encounter in diverse contexts, and when relevant. The latter component deals

[1] Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia; anuska.ferligoj@fdv.uni-lj.si

with their ability to discuss or communicate their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions on the implications of this information, or their concerns regarding the acceptability of given conclusions.

Ben-Zvi and Garfield (2004) distinguish between statistical literacy, statistical reasoning, and statistical thinking. They point out that statistical literacy provides the foundation for reasoning and thinking: basic statistical knowledge makes it possible to reason with statistical ideas and to make sense of statistical information. For Ben-Zvi and Garfield, statistical literacy involves understanding and using the basic language and tools of statistics, while statistical reasoning is the way people reason with statistical ideas and make sense of statistical information. On the other side, statistical thinking involves a higher order of thinking than statistical reasoning. Ben-Zvi and Garfield view statistical thinking as the normative use of statistical models, methods and applications in considering or solving statistical problems. It is the way professional statisticians think.

Statistical literacy was also defined by the *W.M. Keck Statistical Literacy Project* as: (1) critical thinking about numbers, about statistics used as evidence in arguments; (2) the ability to read and interpret numbers in statements, surveys, tables and graphs; and (3) the study of how statistical associations are used as evidence of causal connections (Mittag, 2010).

Several attempts have been made to measure statistical literacy. A very complex study by Watson and Callingham (2003) assumed that statistical literacy is a hierarchical construct. Their analysis of a large archival database of over 3000 school students using Rasch analysis supported the hypothesis of a unidimensional construct and suggested six levels of understanding: (1) *idiosyncratic* engagement with context, tautological use of terminology and basic skills associated with one-to-one counting and reading cell values and tables; (2) *informal* engagement with context often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph, and chance calculations; (3) *inconsistent* engagement with context, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas; (4) *consistent non-critical* engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities, and graph characteristics; (5) *critical*, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, qualitative interpretation of chance, and appreciation of variation; and (6) *critical mathematical* engagement with context, using proportional reasoning particularly in chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language.

# 3 Current initiatives

Many (international) institutions, e.g., the International Association for Statistical Education (IASE), the United Nations Economic Commission for Europe (UNECE), educational institutions, statistical offices, statistical societies and associations, are dealing with how to improve statistical literacy. The most important is the *International Statistical Lit-*

*eracy Project* (ISLP, http://iase-web.org/islp/About.php), being carried out by the International Association for Statistical Education (IASE), the education section of the International Statistical Institute (ISI). It is the only international organisation whose focus is to promote national programmes and strives to increase the statistical literacy of all members of society. This project has concentrated on advancing statistical literacy among secondary school-age students via several activities, including a statistical literacy competition with the aim of bringing the use and understanding of statistics into teaching in a natural way. With various resources and activities assisted by international experts, the ISLP is running a very successful campaign across the continents. In recent years, the ISLP aim has been to extend teaching of statistical literacy to other spheres of life as well. The main target groups defined are as follows: citizens and the media, educational institutions (secondary school and upper secondary school-age students), universities and research institutions, decision-makers, libraries, and national statistical agencies. The perspective of the last target group is "how to bring promotion of statistical literacy more visibly on the agenda of all national statistical agencies". (See the Strategy Project of the ISLP: http://iase-web.org/islp/Activities.php?p=Strategy_Project.)

The United Nations Economic Commission for Europe (UNECE) has taken the notion of statistical literacy as the subject of its fourth guide to making data meaningful. As part of the work programme of the Conference of European Statisticians, a Steering Group on Statistical Dissemination and Communication organises annual Work Sessions that are supported by the UNECE Secretariat. The Steering Group aims to promote good practices in statistical organisations' dissemination and communication of information. The last Work Session on the Communication of Statistics was held in Geneva on 18-20 June 2014. In his talk on *Enabling your Stakeholders to use Statistics*, the keynote speaker Georges-Simon Ulrich, director general of the Swiss Federal Statistical Office, highlighted the importance of having a user-driven approach in National Statistical Offices (NSOs) to identify stakeholders with the goal to help them understand the ways the data are collected and to help them use statistics properly. He also stressed the importance of educating journalists and the public about statistics by improving statistical literacy. Four sessions were organised: Statistical literacy; Communication with respondents and evaluation of communication campaigns; Quick wins on low and zero budgets; and Good practices in electronic publications. Representatives of different countries presented their efforts to improve statistical literacy in their countries in the session on statistical literacy. The Austrian representative presented a project to improve statistical literacy in schools. The Canadian representative talked about a strategy that combines traditional and innovative communication practices to improve the accuracy of media coverage of the Canadian economy and society. The U.S. representatives presented the international programme *Census at School* that provides educators, students and families with an understanding of the relevance and importance of the Census. A presentation on behalf of the U.S. Census Bureau was about a variety of information tools and presentation methods in their communication efforts. The key motivation is to be able to meet the needs of the many audiences that consume statistical information. The need to constantly adapt to rapidly changing technology to deliver information was stressed. The representative from the Netherlands presented efforts and strategies for how to reach young users of statistics. The Latvian representative presented the project called *School Corner* to promote the

understanding of statistical data and proper usage of data among pupils in secondary schools. In addition, in the other three sessions some important issues for improving statistical literacy were discussed.

Several other international and national projects are also related to statistical literacy. One of these is the *W.M. Keck Statistical Literacy Project* at Augsburg College in the U.S.A. The project's primary goal is to present statistical literacy as an interdisciplinary activity. As such, it has overlaps with quantitative reasoning, quantitative literacy, numeracy, and statistical reasoning. A second goal is to present statistical literacy as the study of statisticians in everyday arguments. Milo Schield directs the project and is also the webmaster of Statistical Literacy on the Web (`www.StatLit.org`), a key site for articles, books and activities related to statistical literacy.

Two established projects, *CensusAtSchool* and *ExperimentsAtSchool*, collect data from school-aged learners and this motivates students to analyse the data, and teachers to teach statistics in an exciting way.

A EU-funded project seeking to promote statistical literacy amongst young people by providing an innovative e-course involved the National Statistical Offices of Malta and Finland and the University of Hagen in Germany. The primary goal was to encourage international cooperation between statistical agencies and educational institutions to promote statistical literacy (Mittag, 2010). The main result of this project is a freely available *eCourse in Statistics* with the clear objective of fostering statistical literacy among both the local and international community of statistical users (`http://www.fernuni-hagen.de/statliteracy/`).

Many presentations at the International Conferences on Teaching Statistics (ICOTS) deal with statistical literacy. At the ICOTS 2014 conference, a special topic on *Statistical Literacy in Wider Society* was organised, featuring many presentations. The goal of the topic was "to develop sustainable initiatives which enable citizens to lead and extend debates, in the media and elsewhere, on issues of inequality, crime, effects of smoking, use of alcohol, and support for societal preferences. This democratic imperative leads us to questions such as: How can we encourage people to want to engage in statistical learning? How can we contribute to subject-specific learning of relevant statistical knowledge? How do we enrich our understanding of statistical literacy and methods by which it can be attained and sustained? These invited sessions seek to explore and enrich a variety of effective practices and interventions" (`http://icots.info/9/topic.php?t=7`).

In addition, statistical societies are active in promoting the public understanding of statistics. The Royal Statistical Society launched a ten-year statistical literacy campaign in 2010. Several statistical societies are giving awards to journalists for their correct use of statistical data.

In the last few decades, there has been a greater number of publications dealing with statistical literacy issues. These include the following. In 2004 Ben-Zvi and Garfield edited a very important book dealing with statistical literacy *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. The editors' aim was to provide a useful resource for educators and researchers interested in helping students at all educational levels to develop statistical literacy. The meetings leading up to the book were the Fifth International Conference on Teaching Statistics (ICOTS-5) in 1998, the first, second and third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL_1,2,3) held in 1999, 2001 and 2003 respectively. Selected papers are included

in the resulting book. The first chapter provides basic definitions of statistical literacy, reasoning and thinking. The next four chapters provide an overview of these topics from historical, psychological, and educational perspectives. In the following chapters, particular types of statistical reasoning with key practical implications related to instruction, assessment and research are discussed. The last chapter describes the current state of statistics education research, and implications for teaching statistics.

The former director of the ISLP, Juana Sanchez edited the book *Government Statistical Offices and Statistical Literacy* in 2008. The book's objective was "to explain the process to the emergence of successful, currently active, programs of NSOs to educate the general public in statistics". The editor invited authors committed to improving of statistical literacy in their countries and who had been managing the programmes for a long time, meaning that these programmes reached the front page of the NSOs' web-pages and thereby outside learners and the general public. Authors from NSOs from Portugal, New Zealand, Italy, Finland, Australia and Canada were invited to present good practices for improving statistical literacy in their countries. The goal of the editor and the authors was to encourage other NSOs that had discontinued or never even started programmes on statistical literacy.

A special issue on statistical literacy was published in the *Statistical Journal of the IAOS* in 2011. Beside papers presenting innovative teaching methods to improve statistical literacy, it included papers dealing with innovative work programmes and initiatives in NSOs to improve statistical literacy and the statistical literacy need of different segments of the community. Forbes et al. describe how Statistics New Zealand provides products that support the interpretation of the data they produce for schools, universities and the general public. Townsend argues that educating the public about the world of data can help raise the profile of NSOs in the public mind. Statistics Canada provides outreach and resources designed to improve statistical literacy, working with teachers and students of statistics. Helenius and Mikkela argue that NSOs need to promote statistical literacy in society, with the added consequence of maintaining the legitimacy of NSOs in society. They provide good practices of co-operation between an NSO and various user groups (e.g., the media, educational institutes, members of parliament and citizens). Sanchez et al. argue that NSOs must become more involved in the promotion of statistical literacy, and work together with national statistical societies, international organisations as the ISLP, and national educational institutions stakeholders that share an interest in promoting statistical literacy in different segments of society. Similarly, Gal and Murray pointed out that "improving the effectiveness of information products and services created by statistics agencies requires awareness of four general issues: the factors that affect the difficulty of finding and comprehending statistical products and services, the nature of clients' statistical literacy, the existence of individual or group differences in statistical literacy; and the information needs of different customer groups".

# 4 Who can contribute to better statistical literacy, and how?

As mentioned, several actors are able to contribute to better statistical literacy, e.g., educational institutions, statistical offices, statistical associations, and the media. The more these various actors cooperate in the efforts to improve statistical literacy, the better the results. In the above overview of the current initiatives and main publications seeking to improve statistical literacy, many ideas and good practices were mentioned. Here, some recommendations based on these ideas are suggested for different actors.

## 4.1 Educational institutions

Ruth Carver, the president of the American Statistical Association, pointed out in her presidential message entitled *Statistical Literacy and the 2013 International Year of Statistics* that "statistical literacy can no longer be viewed as a skill needed by a select few; it is essential knowledge required by all that must be developed beginning at an early age and built on throughout one's school years", and later in her message "to reach the goal of a statistically literate citizenry, it is crucial for teachers at all levels to be statistically literate themselves and to possess the pedagogical tools necessary to provide quality learning experiences that develop and deepen their students' statistical understanding" (`http://www.statlit.org/pdf/2012-ASA-Presidents-Message-Statistical-Literacy.pdf`). It is essential that properly educated statisticians teach statistics at all educational levels, from elementary to doctoral. There are still some European countries that have no university programmes on statistics at any level. Also in Slovenia there is no undergraduate programme on statistics, although such education at the master's and doctoral level has been provided for the last 12 years. At least in the European case, some help from European statistical institutions would be appreciated in establishing appropriate statistical programmes. Special attention should also be paid to master's and doctoral programmes on official statistics that target a very important segment of professional government statisticians whose work is essential for policy decision-makers and the segment of professional positions in a wide range of organisations and companies conducting large-scale statistical work.

A lot of effort to impart statistical knowledge to improve statistical literacy on lower levels of education is entailed in different international projects as previously mentioned (e.g., ISLP, *CensusAtSchool*, *ExperimentsAtSchool*). It is important that representatives (individuals or institutions) of as many countries as possible collaborate in these projects. It is especially important to use these very good, internationally developed tools to improve statistical literacy at lower education levels, although they are unfortunately frequently only published in the English language. Therefore, the results of these projects should be appropriately translated into their own languages and promoted to the educational institutions in their countries.

Educational institutions (together with statistical institutions and statistical associations) have to organise meetings, seminars and public discussions for statistics teachers at all education levels to harmonise the statistical terminology used, to logically link statistical topics at different levels of education etc. It is crucial that the individuals and

institutions preparing and implementing statistical courses and programmes at all education levels (from elementary school to doctoral programmes) cooperate with each other. It is important to establish organisational ways to control whether there are problems in the statistical education in the education system at all levels, e.g., if certain topics in statistics are missing on a particular level, too much overlap within and between levels, too difficult for a given level of students, a lack of connections with the data producers, or if the lecturers are appropriately educated.

## 4.2 Statistical Offices and other statistical institutions

It is encouraging that in the last decade National Statistical Offices (NSOs) have shown their stronger awareness of statistical literacy. As mentioned above, several initiatives for improving statistical literacy have been proposed by some NSOs that can also be implemented by the other NSOs. Statistical institutions can transmit the statistical knowledge about the data they collect, the methodologies they use, possible methods to analyse the data, typical abuses or misunderstandings of statistical concepts and data etc. to different segments of users (e.g., business enterprises, governmental sector, researchers, students, the general public, journalists) directly or via the media.

The direct approaches could include:

- the publication of brief and easy-to-read information on the most visible webpages (e.g., webpages of NSOs, statistical associations, educational institutions that organise programmes on statistics) and in different media concerning selected statistical methodologies or data or statistical activities (e.g., information on data sources, why and how they were obtained);

- advising different segments of the population separately about the proper use and interpretation of statistical data;

- organising seminars for different segments of users of statistical data on selected statistical topics for a better understanding of statistical results; and

- the presentation of typical abuses or misunderstandings of statistical concepts and data.

Very important communication with different population segments, especially the adult population, to improve statistical literacy can also be achieved with the help of traditional and especially new media (e.g., the Internet and social media). As Gabrielle Beaudoin stressed at the last Work Session on the Communication of Statistics organised by the Conference of European Statisticians (UNECE), "... Statistics Canada has adopted a strategy that combines traditional and innovative communication practices, with the goal of expanding coverage and improving the accuracy of media coverage about the Canadian economy and society" and later "The proactive, multi-channel approach (was) adopted by the agency to educate journalists and be more responsive to their needs. Media relations activities span from determining the content and style of statistical releases, to hosting concept brief sessions and media lockups, up to training spokespersons and publishing new media content to increase Canadians' understanding of the state of the country".

Several other NSOs have developed a similar good relationships with the media (e.g., Statistics New Zealand, see Harraway and Forbes, 2013), but many of them still do not use the media enough to promote statistical literacy to different segments of the population. Of course, the experience of Statistics Canada with the media is very useful and is recommended to be followed by other statistical institutions, especially NSOs.

As mentioned, journalists have to be properly statistically educated and, to achieve this, there must be cooperation of NSOs and other statistical institutions, educational institutions and statistical associations with the media. There are many ways for achieving this goal. One possibility is to take advantage of an NSO's regular meetings at the end of each month with the media where they present information on the socio-economic indicators of the country. These meetings could be used for a short 'educational' purpose to ensure a proper understanding of the statistical concepts and data, along with examples of misunderstanding. Linking NSOs and other authorised institutions with the users of statistical data could also lead to a better understanding of which data are needed and facilitate searching for better solutions for the planning of statistical data collection, processing and publication. A good example of such cooperation with users was established at the Statistical Office of the Republic of Slovenia (SORS) many years ago. It established 23 statistical advisory committees with around 400 non-SORS members and about 100 SORS members for individual fields of national statistics where experts from different users institutions together with representatives of the statistical office put in a lot of effort to discuss which data are missing, which are no longer relevant, about how to provide quality, timely and relevant statistics.

### 4.3 Statistical societies

The main task of a statistical society is to link various actors (e.g., educational institutions, statistical offices, and the media) and to provide more harmonised efforts for improving statistical literacy. Many national statistical societies are providing awards for good practices of the correct transmission/publication of statistical concepts and data, e.g., awarding journalists for helping the statistical community to improve statistical literacy among different population segments. This practice is also recommended to those societies that have not yet introduced it.

## 5 Conclusion

Statistical literacy has received growing attention in the last decades. In earlier decades, most of the work on statistical education was done in primary and secondary schools with a focus on statistical literacy, statistical reasoning, and statistical thinking. Yet only in the last decade NSOs' awareness of statistical literacy has been present. Some NSOs have shown a broader responsibility than the mere production of data. They are increasingly proactive in improving statistical literacy and making data more accessible. There are also several actions by other statistical institutions, statistical societies and education institutions to improve statistical literacy but a lot remains to be done in the future to improve the statistical literacy of different segments of the population. The cooperation of

all actors dealing with statistics, especially between NSOs and academics, is important for hastening the process of improving statistical literacy.

## Acknowledgements

This paper was presented as an invited paper at the Conference of European Statistics Stakeholders (November 2014, Rome). The author benefited greatly from the discussions with Slovenian colleagues Irena Križman, Metka Zaletel, Tomaž Smrekar and Franc Mali, and with the members of the European Statistical Advisory Committee (ESAC), especially with Denise Lievesley, Irena Kotowska, Ladislav Kabát and Emilia Titan. This research has been supported by the Slovenian Research Agency (P5-0168).

## References

[1] Beaudoin, G. (2014): Meeting the information needs of news media to increase citizens' understanding of statistical findings. Paper presented at *Work Session on the Communication of Statistics*. UNECE.

[2] Ben-Zvi, D., Garfield, J. (2004): Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In Ben-Zvi, D and Garfield, J (eds.): *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pp. 3-15. Kluwer, Boston.

[3] Ben-Zvi, D., Garfield, J. (eds.) (2004): *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Kluwer, Boston.

[4] Forbes, S., Cameden, M.. Pihama, N., Bucknall, P., Pfannkuch, M. (2011): Official statistics and statistical literacy: They need each other. *Statistical Journal of the IAOS*, **27**, 113–128.

[5] Gal, I. (2002): Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, **70**, 1–25.

[6] Gal, I., Scott T., Murray, S.T. (2011): Responding to diversity in users' statistical literacy and information needs: Institutional and educational implications. *Statistical Journal of the IAOS*, **27**, 185–195.

[7] Harraway, J.A., Forbes, S.D. (2013): Partnership between national statistics offices and academics to increase official statistical literacy. *Statistical Journal of the IAOS*, **29**, 31–40.

[8] Helenius, R., Mikkela, H. (2011): Statistical literacy and awareness as strategic success factors of a national statistical office – the case of Statistics Finland. *Statistical Journal of the IAOS*, **27**, 137–144.

[9] Lancaster, G.A. (2011): How statistical literacy, official statistics and selfdirected learning shaped social enquiry in the 19th and early 20th centuries. *Statistical Journal of the IAOS*, **27**, 99–111.

[10] Mittag, H.-J. (2010): Promoting statistical literacy: A European pilot project to bring official statistics into university and secondary school classrooms. Invited paper at *ICOTS-8*. Available at
http://www.fernuni-hagen.de/jmittag/publikationnen/ICOTS8-2010.pdf. Cited 1 Dec 2015

[11] Sanchez, J. (ed.) (2010): Government Statistical Offices and Statistical Literacy. *ISLP*. Available at
http://www.stat.ucla.edu/~jsanchez/books/Front-matter-foreword.pdf. Cited 1 Dec 2015

[12] Sanchez, J., Forbes, S., Campos, P., Giacche, P., Townsend, M., Mooney, G., Helenius, R. (2011): The millennium development goals, national statistical offices, the international statistical literacy project and statistical literacy in schools. *Statistical Journal of the IAOS*, **27**, 157–171.

[13] Schield, M. (2004): Statistical literacy and liberal education at Augsburg College. Peer Review Summer Issue, Assoc. of American Colleges and Universities. Available at
http://web.augsburg.edu/~schield/milopapers/2004schieldaacu.pdf. Cited 1 Dec 2015

[14] Townsend, M. (2011): The national statistical agency as educator. *Statistical Journal of the IAOS*, **27**, 129–136.

[15] Ulrich, G.-S. (2014): Enabling your stakeholders to use statistics. Keynote presentation at *Work Session on the Communication of Statistics*. UNECE.

[16] Wallman, K.K. (1993): Enhancing Statistical Literacy: Enriching Our Society. *JASA*, **88**, 1–8.

[17] Watson, J., Callingham, R. (2003): Statistical literacy: A complex hierarchical construct. Available at
http://www.researchgate.net/profile/Jane_Watson2/publication/247643173_STATISTICAL_LITERACY_A_COMPLEX_HIERARCHICAL_CONSTRUCT1/links/0deec529e5421c83a6000000.pdf. Cited 1 Dec 2015

Metodološki zvezki, Vol. 12, No. 1, 2015, 11-24

# Asymptotic Normality of the Optimal Solution in Multiresponse Surface Mathematical Programming

José A. Díaz-García[1]
Francisco J. Caro-Lopera [2]

**Abstract**

An explicit form for the perturbation effect on the matrix of regression coefficients on the optimal solution in multiresponse surface methodology is obtained in this paper. Then, the sensitivity analysis of the optimal solution is studied and the critical point characterisation of the convex program, associated with the optimum of a multiresponse surface, is also analysed. Finally, the asymptotic normality of the optimal solution is derived by the standard methods.

## 1 Introduction

The multiesponse surface methodology explores the relationships among several explanatory variables and more than one response variables. The addressed methodology considers a sequence of designed experiments in order to obtain a simultaneous optimal response. To reach this aim the method uses a second-degree polynomial model for each response variable. With that constraint the technique is just an approximation, but they are succeed in literature because such models can be easily interpreted, estimated and applied; moreover they perform well under the usual uncertainty about the process or phenomenon under consideration. In fact, a number of laws in sciences are usually explained with second-degree polynomial models, given that the first and second corresponding flows are well understood and they explain some crucial intrinsic property of the phenomenon.

In recent decades, the multiresponse surface methodology has attracted the attention of many quality engineers in different industries. Quality improvement or optimization for such process or phenomenons, needs precise identification of the operation stages and their effects on the response variables. Therefore, multistage systems require special methods and solutions, since applying uni-response surface techniques may lead to suboptimal or even inaccurate results. Some examples of the mentioned industries are: agricultural, pharmaceutical, chemical, assembly, semiconductor, textile, and petroleum industries as well as queuing, healthcare, traffic control, and transportation systems.

Start by assuming that a researcher knows a process or phenomenon and a corresponding set of observable responses variables $Y_1, \cdots, Y_r$ which depends on some input

[1]Universidad Autónoma Agraria Antonio Narro, Calzada Antonio Narro 1923, Col. Buenavista, 25315 Saltillo, Coahuila, Mexico, jadiaz@uaaan.mx

[2] Department of Basic Sciences, Universidad de Medellín, Medellín, Colombia, fjcaro@udem.edu.co

variables, $x_1, \ldots x_n$. Suppose also that the input variables $x_{i's}$ can be controlled by the researcher with a minimum error.

Typically, we have that

$$Y_k(\mathbf{x}) = \eta_k(x_1, \ldots x_n), \quad k = 1, \ldots, r, \text{ and } \mathbf{x} = (x_1, \ldots x_n)', \tag{1.1}$$

where the form of the functions $\eta_k(\cdot)$'s, usually termed as the true response surface, are unknown and perhaps, very complex. The success of the response surface methodology depends on the approximation of $\eta_k(\cdot)$ by a polynomial of low degree in some particular region.

As it was mentioned, in the context of this paper we will assume that $\eta_k(\cdot)$ can be soundly approximated by a polynomial of second order, that is

$$Y_k(\mathbf{x}) = \beta_{0k} + \sum_{i=1}^{n} \beta_{ik} x_i + \sum_{i=1}^{n} \beta_{iik} x_i^2 + \sum_{i=1}^{n} \sum_{j>i}^{n} \beta_{ijk} x_i x_j \tag{1.2}$$

where the unknown parameters $\beta_{j's}$ can be estimated via regression's techniques, as it shall be described in the subsequent section.

We also study the levels of the input variables $x_{i's}$ such that the response variables $Y_1, \cdots, Y_r$ are simultaneously minimal (optimal). This can be achieved if the following multiobjetive mathematical program is solved

$$\begin{array}{c} \min\limits_{\mathbf{x}} \begin{pmatrix} Y_1(\mathbf{x}) \\ Y_2(\mathbf{x}) \\ \vdots \\ Y_r(\mathbf{x}) \end{pmatrix} \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X}, \end{array} \tag{1.3}$$

where $\mathfrak{X}$ is certain operating region for the input variables $x_{i's}$.

Now, two questions, closely related, can be observed:

1. When the estimations of (1.2) for $k = 1, \ldots, r$ are considered into (1.3), the critical point $\mathbf{x}^*$ obtained as solution shall be a function of the estimators $\widehat{\beta}_{j's}$ of the $\beta_{j's}$. Thus, given that$\widehat{\beta}_{j's}$ are random variables, then $\mathbf{x}^* \equiv \mathbf{x}^*(\widehat{\beta}_{j's})$ is a random vector too. So, under the assumption that the distribution of $\widehat{\boldsymbol{\beta}}$ is known, then, we ask for the distribution of $\mathbf{x}^*(\widehat{\beta}_{j's})$.

2. And, given that a point estimate of $\mathbf{x}^*(\widehat{\beta}_{j's})$ should not be sufficient, then we would ask also for an estimated region or an estimated interval.

In particular, the distribution of the critical point in a univariate response surface model was studied by (Díaz García and Ramos-Quiroga, 2001, 2002), when $y(\mathbf{x})$ is defined as an hyperplane.

Now, in the context of the mathematical programming problems, the sensitivity analysis studies the effect of small perturbations in: (1) the parameters on the optimal objective

function value and (2) the critical point. In general, these parameters shape the objective function and constraint the approach to the mathematical programming problem. In particular, (Jagannathan, 1977; Dupačová, 1984; Fiacco and Ghaemi, 1982) have studied the sensitivity analysis of the mathematical programming, among many other authors. As an immediate consequence of the sensitivity analysis, the corresponding asymptotic normality study of the critical point emerges naturally, which can be performed by standard methods of mathematical statistics (see similar results for the case of maximum likelihood estimates in (Aitchison and Silvey, 1958)). This last consequence makes the sensitivity analysis as an interesting source of statistical research. However, this approach must be fitted into the classical philosophy of the sensitivity analysis; i.e., we need to translate the general sensitivity analysis methodology into the statistical language. This involves, for example, to study the effect on the model estimators of adding and/or excluding variables and/or observations, see (Chatterjee and Hadi, 1988).

This paper proposes a solution in order to establish the effect of perturbations of the matrix of regression parameters on the optimal solution of the multiresponse surface model and the asymptotic normality of the critical point. First, in Section 2 some notation is proposed. Then, the multiresponse surface mathematical program is set in Section 3 as a multiobjective mathematical programming problem and a general solution is considered in terms of a functional. Then, the characterisation of the critical point is given in Section 4 by stating the first-order and second-order Kuhn-Tucker conditions. Finally, the asymptotic normality of a critical point is established in Section 5 and for a particular form of the functional, the asymptotic normality of a critical point is also derived.

# 2 Notation

For the sake of completeness, the main properties and usual notations are given here. But for a detailed discussion of the multiresponse surface methodology we recommend references (Khuri and Cornell, 1987; Khuri and Conlon, 1981, Chap. 7).

Let $N$ be the number of experimental runs and $r$ be the number of response variables, which can be measured for each setting of a group of $n$ coded variables $x_1, x_2, \ldots, x_n$. We assume that the response variables can be modeled by a second order polynomial regression model in terms of $x_i$, $i = 1, \ldots, n$. Hence, the $k^{th}$ response model can be written as

$$\mathbf{Y}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \quad k = 1, \ldots, r, \tag{2.1}$$

where $\mathbf{Y}_k$ is an $N \times 1$ vector of observations on the $k^{th}$ response, $\mathbf{X}_k$ is an $N \times p$ matrix of rank $p$ termed the design or regression matrix, $p = 1 + n + n(n+1)/2$, $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of unknown constant parameters, and $\boldsymbol{\varepsilon}_k$ is a random error vector associated with the $k^{th}$ response. For purposes of this study, it is assumed that $\mathbf{X}_1 = \cdots = \mathbf{X}_r = \mathbf{X}$. Therefore, (2.1) can be written as

$$\mathbf{Y} = \mathbf{X}\mathbb{B} + \mathbb{E} \tag{2.2}$$

where $\mathbf{Y} = \left[\mathbf{Y}_1 \vdots \mathbf{Y}_2 \vdots \cdots \vdots \mathbf{Y}_r\right]$, $\mathbb{B} = \left[\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2 \vdots \cdots \vdots \boldsymbol{\beta}_r\right]$, moreover

$$\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{nk}, \beta_{11k}, \ldots, \beta_{nnk}, \beta_{12k}, \ldots, \beta_{(n-1)nk})'$$

and $\mathbb{E} = \left[\boldsymbol{\varepsilon}_1 \vdots \boldsymbol{\varepsilon}_2 \vdots \cdots \vdots \boldsymbol{\varepsilon}_r\right]$, such that $\mathbb{E} \sim \mathcal{N}_{N\times r}(\mathbf{0}, \mathbf{I}_N \otimes \boldsymbol{\Sigma})$ i.e. $\mathbb{E}$ has an $N \times r$ matrix multivariate normal distribution with $\mathrm{E}(\mathbb{E}) = \mathbf{0}$ and $\mathrm{Cov}(\mathrm{vec}\,\mathbb{E}') = \mathbf{I}_N \otimes \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a $r \times r$ positive definite matrix. Now, if $\mathbf{A} = \left[\mathbf{A}_1 \vdots \mathbf{A}_2 \vdots \cdots \vdots \mathbf{A}_r\right]$, with $\mathbf{A}_j$, $j = 1, \cdots, r$ the columns of $\mathbf{A}$; then $\mathrm{vec}\,\mathbf{A} = (\mathbf{A}_1', \mathbf{A}_2', \ldots, \mathbf{A}_r')'$ and $\otimes$ denotes the direct (or Kronecker) product of matrices, see (Muirhead, 1982, Theorem 3.2.2, p. 79). In addition denote

- $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$: The vector of controllable variables or factors.
- $\widehat{\mathbb{B}} = \left[\widehat{\boldsymbol{\beta}}_1 \vdots \widehat{\boldsymbol{\beta}}_2 \vdots \cdots \vdots \widehat{\boldsymbol{\beta}}_r\right]$: The least squares estimator of $\mathbb{B}$ given by

  $$\widehat{\mathbb{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

  from where

  $$\widehat{\boldsymbol{\beta}}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_k = (\widehat{\beta}_{0k}, \widehat{\beta}_{1k}, \ldots, \widehat{\beta}_n, \widehat{\beta}_{11k}, \ldots, \widehat{\beta}_{nnk}, \widehat{\beta}_{12k}, \ldots, \widehat{\beta}_{(n-1)nk})'$$

  $k = 1, 2, \ldots, r$. Moreover, under the assumption that $\mathbb{E} \sim \mathcal{N}_{N\times r}(\mathbf{0}, \mathbf{I}_N \otimes \boldsymbol{\Sigma})$, then $\widehat{\mathbb{B}} \sim \mathcal{N}_{p\times r}(\mathbb{B}, (\mathbf{X}'\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma})$, with $\mathrm{Cov}(\mathrm{vec}\,\widehat{\mathbb{B}}') = (\mathbf{X}'\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}$.
- $\mathbf{z}(\mathbf{x}) = (1, x_1, x_2, \ldots, x_n, x_1^2, x_2^2, \ldots, x_n^2, x_1x_2, x_1x_3 \ldots, x_{n-1}x_n)'$.
- $\widehat{\boldsymbol{\beta}}_{1k} = (\widehat{\beta}_{1k}, \ldots, \widehat{\beta}_{nk})'$ and

  $$\widehat{\mathbf{B}}_k = \frac{1}{2}\begin{pmatrix} 2\widehat{\beta}_{11k} & \widehat{\beta}_{12k} & \cdots & \widehat{\beta}_{1nk} \\ \widehat{\beta}_{21k} & 2\widehat{\beta}_{22k} & \cdots & \widehat{\beta}_{2nk} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\beta}_{n1k} & \widehat{\beta}_{n2k} & \cdots & 2\widehat{\beta}_{nnk} \end{pmatrix}$$

- $$\begin{aligned} \widehat{Y}_k(\mathbf{x}) &= \mathbf{z}'(\mathbf{x})\widehat{\boldsymbol{\beta}}_k \\ &= \widehat{\beta}_{0k} + \sum_{i=1}^{n}\widehat{\beta}_{ik}x_i + \sum_{i=1}^{n}\widehat{\beta}_{iik}x_i^2 + \sum_{i=1}^{n}\sum_{j>i}^{n}\widehat{\beta}_{ijk}x_ix_j \\ &= \widehat{\beta}_{0k} + \widehat{\boldsymbol{\beta}}_{1k}'\mathbf{x} + \mathbf{x}'\widehat{\mathbf{B}}_k\mathbf{x} : \end{aligned}$$

  The response surface or predictor equation at the point $\mathbf{x}$ for the k$^{th}$ response variable. For the aim of this paper it is assumed that $\widehat{\mathbf{B}}_k$, $k = 1, \ldots, r$, are positive definite matrices. Note that this last assumption is not always true and should be verified, for example via the canonical analysis, see (Khuri and Cornell, 1987, Subsection 5.5.1, pp. 180-186). [2]
- $\widehat{\mathbf{Y}}(\mathbf{x}) = \left(\widehat{Y}_1(\mathbf{x}), \widehat{Y}_2(\mathbf{x}), \ldots, \widehat{Y}_r(\mathbf{x})\right)' = \widehat{\mathbb{B}}'\mathbf{z}(\mathbf{x})$: The multiresponse surface or predicted response vector at the point $\mathbf{x}$.

[2] Observe that, alternatively can be assumed that $\widehat{\mathbf{B}}_k$, $k = 1, \ldots, r$, are negative definite matrices and then in equation (3.1) the minimization should be replaced by maximization.

- $\widehat{\mathbf{\Sigma}} = \dfrac{\mathbf{Y}'(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}}{N-p}$: The estimator of the variance-covariance matrix $\mathbf{\Sigma}$ such that $(N-p)\widehat{\mathbf{\Sigma}}$ has a Wishart distribution with $(N-p)$ degrees of freedom and the parameter $\mathbf{\Sigma}$; this fact is denoted as $(N-p)\widehat{\mathbf{\Sigma}} \sim \mathcal{W}_r(N-p, \mathbf{\Sigma})$. Here, $\mathbf{I}_m$ denotes an identity matrix of order $m$.

- Finally, notice that
$$E(\widehat{\mathbf{Y}}(\mathbf{x})) = E(\widehat{\mathbb{B}}'\mathbf{z}(\mathbf{x})) = \mathbb{B}'\mathbf{z}(\mathbf{x}) \tag{2.3}$$
and
$$\mathrm{Cov}(\widehat{\mathbf{Y}}(\mathbf{x})) = \mathbf{z}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}(\mathbf{x})\mathbf{\Sigma}. \tag{2.4}$$
An unbiased estimator of $\mathrm{Cov}(\widehat{\mathbf{Y}}(\mathbf{x}))$ is given by
$$\widehat{\mathrm{Cov}}(\widehat{\mathbf{Y}}(\mathbf{x})) = \mathbf{z}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}(\mathbf{x})\widehat{\mathbf{\Sigma}}. \tag{2.5}$$

# 3 Multiresponse surface mathematical programming

In the following sections, we make use of the multiresponse mathematical programming and multiobjective mathematical programming. For convenience, the concepts and notations required are listed below in terms of the estimated model of multiresponse surface mathematical programming, but further detailed properties can be found in (Khuri and Conlon, 1981; Khuri and Cornell, 1987; Ríos *et al.*, 1989; Steuer, 1986; Miettinen, 1999).

The multiresponse mathematical programming or multiresponse optimisation (MRO) problem is proposed, in general, as follows

$$\begin{array}{c} \min\limits_{\mathbf{x}} \widehat{\mathbf{Y}}(\mathbf{x}) = \min\limits_{\mathbf{x}} \begin{pmatrix} \widehat{Y}_1(\mathbf{x}) \\ \widehat{Y}_2(\mathbf{x}) \\ \vdots \\ \widehat{Y}_r(\mathbf{x}) \end{pmatrix} \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X}. \end{array} \tag{3.1}$$

It is a nonlinear multiobjective mathematical programming problem, see (Steuer, 1986; Ríos *et al.*, 1989; Miettinen, 1999); and $\mathfrak{X}$ denotes the experimental region, usually taken as a hypersphere

$$\mathfrak{X} = \{\mathbf{x} | \mathbf{x}'\mathbf{x} \leq c^2, c \in \Re\},$$

where, $c$ is set according to the experimental design model under consideration, see (Khuri and Cornell, 1987). Alternatively, the experimental region can be taken as a hypercube

$$\mathfrak{X} = \{\mathbf{x} | l_i < x_i < u_i, \quad i = 1, 2, \ldots, n\},$$

where

$$\mathbf{l} = (l_1, l_2, \ldots, l_n)',$$

defines the vector of lower bounds of factors and

$$\mathbf{u} = (u_1, u_2, \ldots, u_n)',$$

gives the vector of upper bounds of factors. Alternatively (3.1) can be written as

$$\min_{\mathbf{x}\in\mathfrak{X}} \widehat{\mathbf{Y}}(\mathbf{x}).$$

In the response surface methodology context, the multiobjective mathematical programs rarely contain a point $\mathbf{x}^*$ which can be considered as an optimum, i.e. few cases satisfy the requirement that $\widehat{Y}_k(\mathbf{x})$ is minimum for all $k = 1, 2, \ldots, r$. From the viewpoint of multiobjective mathematical programming, this justifies the following notion of the *Pareto point*:

> *We say that* $\widehat{\mathbf{Y}}^*(\mathbf{x})$ *is a Pareto point of* $\widehat{\mathbf{Y}}(\mathbf{x})$*, if there is no other point* $\widehat{\mathbf{Y}}^1(\mathbf{x})$ *such that* $\widehat{\mathbf{Y}}^1(\mathbf{x}) \leq \widehat{\mathbf{Y}}^*(\mathbf{x})$*, i.e. for all* $k$*,* $\widehat{Y}_k^1(\mathbf{x}) \leq \widehat{Y}_k^*(\mathbf{x})$ *and* $\widehat{\mathbf{Y}}^1(\mathbf{x}) \neq \widehat{\mathbf{Y}}^*(\mathbf{x})$.

(Steuer, 1986; Ríos *et al.*, 1989; Miettinen, 1999) established the existence criteria for Pareto points in a multiobjective mathematical programming problem and the extension of scalar mathematical programming (*Kuhn-Tucker's conditions*) to the vectorial case.

Methods for solving a multiobjective mathematical program are based on the existing information about a particular problem. There are three possible scenarios: when the investigator possesses either complete, partial or null information, see (Ríos *et al.*, 1989; Miettinen, 1999; Steuer, 1986). In a response surface methodology context, complete information means that the investigator understands the population, in such a way, that it is possible to propose a *value function* reflecting the importance of each response variable. In partial information, the investigator knows deeply the main response variable of the study and this is a sufficient support for the research. Finally, under null information, the researcher only possesses information about the estimators of the response surface parameter, and with this elements, an appropriate solution can be found too.

In general, an approach for solving a multiobjective mathematical program consist of studying an equivalent nonlinear scalar mathematical program, i.e. as a solution of (1.3) is proposed the following problem

$$\begin{array}{c} \min_{\mathbf{x}} f\left(\mathbf{Y}(\mathbf{x})\right) \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X}, \end{array} \tag{3.2}$$

and as a solution of (3.1) is stated the following problem

$$\begin{array}{c} \min_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X}, \end{array} \tag{3.3}$$

where $f(\cdot)$ defines a functional ($f(\cdot)$ is a function that takes functions as its argument, i.e. a function whose domain is a set of functions). Moreover, in the context of multiobjective mathematical programmming, the functional $f(\cdot)$ is such that if $\mathfrak{M} \subset \Re^r$ denotes a set of multiresponse surface functions, then

> *The* ***functional*** *is a function* $f : \mathfrak{M} \to \Re$ *such that* $\min \widehat{\mathbf{Y}}(\mathbf{x}^*) < \min \widehat{\mathbf{Y}}(\mathbf{x}_1) \Leftrightarrow f(\widehat{\mathbf{Y}}(\mathbf{x}^*)) < f(\widehat{\mathbf{Y}}(\mathbf{x}_1)), \quad \mathbf{x}^* \neq \mathbf{x}_1.$

In order to consider a greater number of potential solutions of (3.1), usually studied in the multicriteria mathematical programming, the following alternative problem to (3.3) can be proposed

$$\begin{array}{c} \min\limits_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X} \cap \mathfrak{S}, \end{array} \tag{3.4}$$

where $\mathfrak{S}$ is a subset generated by additional potential constraints, generally derived by a particular technique used for establishing the equivalent scalar mathematical program (3.3). In some particular cases of (3.3), a new fixed parameter may appear, a vector of response weights $\mathbf{w} = (w_1, w_2, \ldots, w_r)'$, and/or a vector of target values for the response vector $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_r)'$. Particular examples of this equivalent univariate objective mathematical programming are the use of goal programming, see (Kazemzadeh *et al.*, 2008), and of the $\epsilon$-constraint model, see (Biles, 1975), among many others. In particular, under the $\epsilon$-constraint model, (3.4) is proposed as

$$\begin{array}{c} \min\limits_{\mathbf{x}} \widehat{Y}_j(\mathbf{x}) \\ \text{subject to} \\ \widehat{Y}_1(\mathbf{x}) \leq \tau_1 \\ \vdots \\ \widehat{Y}_{j-1}(\mathbf{x}) \leq \tau_{j-1} \\ \widehat{Y}_{j+1}(\mathbf{x}) \leq \tau_{j+1} \\ \vdots \\ \widehat{Y}_r(\mathbf{x}) \leq \tau_r \\ \mathbf{x} \in \mathfrak{X}. \end{array} \tag{3.5}$$

# 4 Characterisation of the critical point

In the rest of the paper we shall develop the theory of the problem (3.3); it is easy to see that this problem can be extended with minor modifications to the problem (3.4).

Hereinafter it is assumed that each possible functional $f(\cdot)$ considered in (3.3) is such that the nonlinear scalar mathematical program

$$\begin{array}{c} \min\limits_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) \\ \text{subject to} \\ \mathbf{x} \in \mathfrak{X}, \end{array}$$

defines a convex program.

**Remark 1.** *For example, suppose that such functional $f(\cdot)$ is defined as*

$$\begin{aligned} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) &= \sum_{k=1}^{r} w_k \widehat{Y}_k(\mathbf{x}) \\ &= \sum_{k=1}^{r} w_k \left(\widehat{\beta}_{0k} + \widehat{\boldsymbol{\beta}}_{1k}'\mathbf{x} + \mathbf{x}'\widehat{\mathbf{B}}_k\mathbf{x}\right) \\ &= \sum_{k=1}^{r} w_k\widehat{\beta}_{0k} + \left(\sum_{k=1}^{r} w_k\widehat{\boldsymbol{\beta}}_{1k}'\right)\mathbf{x} + \mathbf{x}'\left(\sum_{k=1}^{r} w_k\widehat{\mathbf{B}}_k\right)\mathbf{x}, \end{aligned}$$

*that is, $f(\cdot)$ is defined as a weight value function, such that, $w_k \geq 0$, $k = 1, \ldots, r$, with $\sum_{k=1}^{r} w_k = 1$, see (Ríos et al., 1989). And observe that, as $\widehat{\mathbf{B}}_k$, $k = 1, \ldots, r$, are positive definite matrices, then $\sum_{k=1}^{r} \delta_k\widehat{\mathbf{B}}_k$ is a positive definite matrix too. Then the equivalent nonlinear scalar mathematical program defined in this manner is a quadratic program, and hence a convex program, see (Rao, 1979, p. 662).* □

Let $\mathbf{x}^*(\widehat{\mathbb{B}}) \in \Re^n$ be the unique optimal solution of program (3.3) with the corresponding Lagrange multiplier $\lambda^*(\widehat{\mathbb{B}}) \in \Re$. The Lagrangian is defined by

$$L(\mathbf{x}, \lambda; \widehat{\mathbb{B}}) = f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) + \lambda(||\mathbf{x}||^2 - c^2). \tag{4.1}$$

Similarly, $\mathbf{x}^*(\mathbb{B}) \in \Re^n$ denotes the unique optimal solution of program (1.3) with the corresponding Lagrange multiplier $\lambda^*(\mathbb{B}) \in \Re$.

Now we establish the local Kuhn-Tucker conditions that guarantee that the Kuhn-Tucker point $\mathbf{r}^*(\widehat{\mathbb{B}}) = \left[\mathbf{x}^*(\widehat{\mathbb{B}}), \lambda^*(\widehat{\mathbb{B}})\right]' \in \Re^{n+1}$ is a unique global minimum of convex program (3.3). First recall that for $f : \Re^n \to \Re$, $\dfrac{\partial f}{\partial \mathbf{x}} \equiv \nabla_{\mathbf{x}}$ denotes the gradient of function $f$.

**Theorem 1.** *The necessary and sufficient conditions that a point $\mathbf{x}^*(\widehat{\mathbb{B}}) \in \Re^n$ for arbitrary fixed $\widehat{\mathbb{B}} \in \Re^p$, be a unique global minimum of the convex program (3.3) is that, $\mathbf{x}^*(\widehat{\mathbb{B}})$ and the corresponding Lagrange multiplier $\lambda^*(\widehat{\mathbb{B}}) \in \Re$, fulfill the Kuhn-Tucker first order conditions*

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda; \widehat{\mathbb{B}}) = \nabla_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) + 2\lambda(\widehat{\mathbb{B}})\mathbf{x} = \mathbf{0} \tag{4.2}$$

$$\nabla_{\lambda} L(\mathbf{x}, \lambda; \widehat{\mathbb{B}}) = ||\mathbf{x}||^2 - c^2 \leq 0 \tag{4.3}$$

$$\lambda(\widehat{\mathbb{B}})(||\mathbf{x}||^2 - c^2) = 0 \tag{4.4}$$

$$\lambda(\widehat{\mathbb{B}}) \geq 0 \tag{4.5}$$

In addition, assume that strict complementarity slackness holds at $\mathbf{x}^*(\mathbb{B})$ with respect to $\lambda^*(\mathbb{B})$, that is

$$\lambda^*(\mathbb{B}) > 0 \Leftrightarrow ||\mathbf{x}||^2 - c^2 = 0. \tag{4.6}$$

Analogously, the Khun-Tucker condition (4.2) to (4.5) for $\widehat{\mathbb{B}} = \mathbb{B}$ are stated next.

**Corollary 1.** *The necessary and sufficient conditions that a point* $\mathbf{x}^*(\mathbb{B}) \in \Re^n$ *for arbitrary fixed* $\mathbb{B} \in \Re^p$*, be a unique global minimum of the convex program (3.2) is that,* $\mathbf{x}^*(\mathbb{B})$ *and the corresponding Lagrange multiplier* $\lambda^*(\mathbb{B}) \in \Re$*, fulfill the Kuhn-Tucker first order conditions*

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}, \lambda; \mathbb{B}) = \nabla_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) + 2\lambda(\mathbb{B})\mathbf{x} &= \mathbf{0} && (4.7)\\ \nabla_{\lambda} L(\mathbf{x}, \lambda; \mathbb{B}) = ||\mathbf{x}||^2 - c^2 &\leq 0 && (4.8)\\ \lambda(\mathbb{B})(||\mathbf{x}||^2 - c^2) &= 0 && (4.9)\\ \lambda(\mathbb{B}) &\geq 0 && (4.10) \end{aligned}$$

*and* $\lambda(\mathbb{B}) = 0$ *when* $||\mathbf{x}||^2 - c^2 < 0$ *at* $[\mathbf{x}^*(\mathbb{B}), \lambda^*(\mathbb{B})]'$.

Observe that, due to the strict convexity of the constraint and objective function, the second-order sufficient condition is evidently fulfilled for the convex program (3.3).

The next result states the existence of a once continuously differentiable solution to program (3.3), see (Fiacco and Ghaemi, 1982).

**Theorem 2.** *Assume that (4.6) holds and the second-order sufficient condition is satisfied by the convex program (3.3). Then*

1. $\mathbf{x}^*(\mathbb{B})$ *is a unique global minimum of program (3.2) and* $\lambda^*(\mathbb{B})$ *is also unique.*

2. *For* $\widehat{\mathbb{B}} \in V_{\varepsilon}(\mathbb{B})$ *(is an* $\varepsilon-$*neighborhood or open ball), there exist a unique once continuously differentiable vector function*

$$\mathbf{r}^*(\widehat{\mathbb{B}}) = \begin{bmatrix} \mathbf{x}^*(\widehat{\mathbb{B}}) \\ \lambda^*(\widehat{\mathbb{B}}) \end{bmatrix} \in \Re^{n+1}$$

   *satisfying the second order sufficient conditions of problem (3.2), such that* $\mathbf{r}^*(\mathbb{B}) = [\mathbf{x}^*(\mathbb{B}), \lambda^*(\mathbb{B})]'$ *and hence,* $\mathbf{x}^*(\widehat{\mathbb{B}})$ *is a unique global minimum of problem (3.3) with associated unique Lagrange multiplier* $\lambda^*(\widehat{\mathbb{B}})$.

3. *For* $\widehat{\mathbb{B}} \in V_{\varepsilon}(\mathbb{B})$*, the status of the constraint is unchanged and* $\lambda^*(\widehat{\mathbb{B}}) > 0 \Leftrightarrow ||\mathbf{x}||^2 - c^2 = 0$ *holds.*

# 5 Asymptotic normality of the critical point

This section considers the statistical and mathematical programming aspects in the sensitivity analysis of the optimum of a estimated multiresponse surface model.

**Theorem 3.** *Assume that:*

1. *For any* $\widehat{\mathbb{B}} \in V_{\varepsilon}(\mathbb{B})$*, the second-order sufficient condition is fulfilled for the convex program (3.3) such that the second order derivatives*

$$\frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \mathbf{x}'}, \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \operatorname{vec}' \widehat{\mathbb{B}}}, \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \lambda}, \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \lambda \partial \mathbf{x}'}, \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \lambda \partial \operatorname{vec} \widehat{\mathbb{B}}}$$

*exist and are continuous in* $\left[\mathbf{x}^*(\widehat{\mathbb{B}}), \lambda^*(\widehat{\mathbb{B}})\right]' \in V_\varepsilon([\mathbf{x}^*(\mathbb{B}), \lambda^*(\mathbb{B})]')$ *and*

$$\frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \mathbf{x}'}$$

*is positive definite.*

2. $\widehat{\mathbb{B}}_\nu$*, the estimator of the true parameter vector* $\mathbb{B}_\nu$*, is based on a sample of size* $N_\nu$ *such that*

$$\sqrt{N_\nu}(\widehat{\mathbb{B}}_\nu - \mathbb{B}_\nu) \sim \mathcal{N}_{p\times r}(\mathbb{B}, \mathbf{\Theta}), \quad \frac{1}{N_\nu}\mathbf{\Theta} = (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}.$$

3. *(4.6) holds for* $\widehat{\mathbb{B}} = \mathbb{B}$*.*

   *Then asymptotically*

$$\sqrt{N_\nu}\left[\mathbf{x}^*(\widehat{\mathbb{B}}) - \mathbf{x}^*(\mathbb{B})\right] \xrightarrow{d} \mathcal{N}_n(\mathbf{0}_n, \mathbf{\Xi}),$$

   *where the* $n \times n$ *variance-covariance matrix*

$$\mathbf{\Xi} = \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) \widehat{\mathbf{\Theta}} \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right)', \quad \frac{1}{N_\nu}\widehat{\mathbf{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1} \otimes \widehat{\mathbf{\Sigma}}$$

   *such that all elements of* $\left(\partial \mathbf{x}^*(\widehat{\mathbb{B}})/\partial \operatorname{vec} \widehat{\mathbb{B}}\right)$ *are continuous on any* $\widehat{\mathbb{B}} \in V_\varepsilon(\mathbb{B})$*; furthermore*

$$\left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) = \left[\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q}(\mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\right]\mathbf{P}^{-1}\mathbf{G},$$

   *where*

$$\mathbf{P} = \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \mathbf{x}'}$$

$$\mathbf{Q} = \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \lambda \partial \mathbf{x}}$$

$$\mathbf{G} = \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \operatorname{vec}' \widehat{\mathbb{B}}}$$

*Proof.* According to Theorem 1 and Corollary 1, the Kuhn-Tucker conditions (4.2)–(4.5) at $\left[\mathbf{x}^*(\widehat{\mathbb{B}}), \lambda^*(\widehat{\mathbb{B}})\right]'$ and the conditions (4.7)–(4.10) at $[\mathbf{x}^*(\mathbb{B}), \lambda^*(\mathbb{B})]'$ are fulfilled for mathematical programs (3.2) and (3.3), respectively. From conditions (4.7)–(4.10) of Corollary 1, the following system equation

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda; \mathbb{B}) = \nabla_{\mathbf{x}} f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) + 2\lambda(\mathbb{B})\mathbf{x} = \mathbf{0} \tag{5.1}$$

$$\nabla_{\lambda} L(\mathbf{x}, \lambda; \mathbb{B}) = ||\mathbf{x}||^2 - c^2 = 0, \tag{5.2}$$

has a solution $\mathbf{x}^*(\mathbb{B}), \lambda^*(\mathbb{B}) > 0, \mathbb{B}$.

The nonsingular Jacobian matrix of the continuously differentiable functions (5.1) and (5.2) with respect to $\mathbf{x}$ and $\lambda$ at $\left[\mathbf{x}^*(\widehat{\mathbb{B}}), \lambda^*(\widehat{\mathbb{B}})\right]'$ is

$$\begin{pmatrix} \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \mathbf{x}\partial \mathbf{x}'} & \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \lambda \partial \mathbf{x}} \\ \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \mathbf{x}'\partial \lambda} & 0 \end{pmatrix}.$$

According to the implicit functions theorem, there is a neighborhood $V_\varepsilon(\mathbb{B})$ such that for arbitrary $\widehat{\mathbb{B}} \in V_\varepsilon(\mathbb{B})$, the system (5.1) and (5.2) has a unique solution $\mathbf{x}^*(\widehat{\mathbb{B}})$, $\lambda^*(\widehat{\mathbb{B}})$, $\widehat{\mathbb{B}}$ and by Theorem 2, the components of $\mathbf{x}^*(\widehat{\mathbb{B}})$, $\lambda^*(\widehat{\mathbb{B}})$ are continuously differentiable function of $\widehat{\mathbb{B}}$, see (Bigelow and Shapiro, 1974). Their derivatives are given by

$$\begin{pmatrix} \dfrac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}} \\ \dfrac{\partial \lambda^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}} \end{pmatrix} = - \begin{pmatrix} \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \mathbf{x}\partial \mathbf{x}'} & \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \lambda \partial \mathbf{x}} \\ \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \mathbf{x}'\partial \lambda} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \dfrac{\partial^2 L(\mathbf{x},\lambda;\widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \operatorname{vec}' \widehat{\mathbb{B}}} \\ 0 \end{pmatrix} \quad (5.3)$$

The explicit form of $(\partial \mathbf{x}^*(\widehat{\mathbb{B}})/\partial \operatorname{vec} \widehat{\mathbb{B}})$ follows from (5.3) and by the formula

$$\begin{pmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} [\mathbf{I} - \mathbf{P}^{-1}\mathbf{Q}(\mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}\mathbf{Q}']\mathbf{P}^{-1} & \mathbf{P}^{-1}\mathbf{Q}(\mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1} \\ (\mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{P}^{-1} & -(\mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1} \end{pmatrix},$$

where $\mathbf{P}$ is symmetric and nonsingular.

Then from assumption 2, (Rao, 1973, (iii), p. 388) and (Bishop *et al.*, 1991, Theorem 14.6-2, p. 493) (see also (Cramér, 1946, p. 353)) we have

$$\sqrt{N_\nu}\left[\mathbf{x}^*(\widehat{\mathbb{B}}) - \mathbf{x}^*(\mathbb{B})\right] \xrightarrow{d} \mathcal{N}_n\left(\mathbf{0}_n, \left(\frac{\partial \mathbf{x}^*(\mathbb{B})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) \boldsymbol{\Theta} \left(\frac{\partial \mathbf{x}^*(\mathbb{B})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right)'\right). \quad (5.4)$$

Finally note that all elements of $(\partial \mathbf{x}^*/\partial \widehat{\mathbb{B}})$ are continuous on $V_\varepsilon(\mathbb{B})$, so that the asymptotical distribution (5.4) can be substituted by

$$\sqrt{N_\nu}\left[\mathbf{x}^*(\widehat{\mathbb{B}}) - \mathbf{x}^*(\mathbb{B})\right] \xrightarrow{d} \mathcal{N}_n\left(\mathbf{0}_n, \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) \widehat{\boldsymbol{\Theta}} \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right)'\right),$$

see (Rao, 1973, (iv), pp.388–389). □

As a particular case, assume that the functional in (3.3) is defined as

$$f\left(\widehat{\mathbf{Y}}(\mathbf{x})\right) = \sum_{k=1}^{r} w_k \widehat{Y}_k(\mathbf{x}), \quad \sum_{k=1}^{r} w_k = 1,$$

with $w_k$ known constants. Then,

**Corollary 2.** *Suppose the hypothesis 1 to 3 of Theorem 3 are fulfilled.*
*Then asymptotically*

$$\sqrt{N_\nu}\left[\mathbf{x}^*(\widehat{\mathbb{B}}) - \mathbf{x}^*(\mathbb{B})\right] \xrightarrow{d} \mathcal{N}_n(\mathbf{0}_n, \boldsymbol{\Xi})$$

*where the $n \times n$ variance-covariance matrix*

$$\boldsymbol{\Xi} = \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) \widehat{\boldsymbol{\Theta}} \left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right)', \quad \frac{1}{N_\nu}\widehat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1} \otimes \widehat{\boldsymbol{\Sigma}}$$

*such that all elements of $\left(\partial \mathbf{x}^*(\widehat{\mathbb{B}})/\partial \operatorname{vec} \widehat{\mathbb{B}}\right)$ are continuous on any $\widehat{\mathbb{B}} \in V_\varepsilon(\mathbb{B})$; furthermore*

$$\left(\frac{\partial \mathbf{x}^*(\widehat{\mathbb{B}})}{\partial \operatorname{vec} \widehat{\mathbb{B}}}\right) = \mathbf{S}^{-1}\left(\frac{\mathbf{x}^*(\widehat{\mathbb{B}})\mathbf{x}^*(\widehat{\mathbb{B}})'\mathbf{S}^{-1}}{\mathbf{x}^*(\widehat{\mathbb{B}})'\mathbf{S}^{-1}\mathbf{x}^*(\widehat{\mathbb{B}})} - \mathbf{I}_n\right)\mathbf{M}\left(\mathbf{x}^*(\widehat{\mathbb{B}})\right),$$

*where*

$$\mathbf{S} = \frac{\partial^2 L(\mathbf{x}, \lambda; \widehat{\mathbb{B}})}{\partial \mathbf{x} \partial \mathbf{x}'} = 2\sum_{k=1}^{r} w_k \widehat{\mathbf{B}}_k - 2\lambda^*(\widehat{\mathbb{B}})\mathbf{I}_n.$$

*and*

$$\begin{aligned}\mathbf{M}(\mathbf{x}) &= \nabla_{\mathbf{x}}\mathbf{z}'(\mathbf{x}) = \frac{\partial \mathbf{z}'(\mathbf{x})}{\partial \mathbf{x}} \\ &= (\mathbf{0}\vdots\mathbf{I}_n\vdots 2\operatorname{diag}(\mathbf{x})\vdots\mathbf{C}_1\vdots\cdots\vdots\mathbf{C}_{n-1}) \in \Re^{n\times p},\end{aligned}$$

*with*

$$\mathbf{C}_i = \begin{pmatrix} \mathbf{0}_1' \\ \vdots \\ \mathbf{0}_{i-1}' \\ \mathbf{x}'\mathbf{A}_i \\ x_i\mathbf{I}_{n-i} \end{pmatrix}, i = 1, \ldots, n-1, \quad \mathbf{0}_j \in \Re^{n-i}, j = 1, \ldots, i-1;$$

*observing that when $i = 1$ (i.e. j = 0), this row does not appear in $\mathbf{C}_1$; and*

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{0}_1' \\ \vdots \\ \mathbf{0}_i' \\ \mathbf{I}_{n-i} \end{pmatrix}, \quad \mathbf{0}_k' \in \Re^{n-i}, k = 1, \ldots, i.$$

*Proof.* The required result follows from Theorem 3 and observing that in this particular case

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda; \mathbb{B}) = \left\{ \begin{array}{l} \mathbf{M}(\mathbf{x})\displaystyle\sum_{k=1}^{r} w_k \boldsymbol{\beta}_k + 2\lambda(\mathbb{B})\mathbf{x} \\ \quad \text{or} \\ \displaystyle\sum_{k=1}^{r} w_k \left[\boldsymbol{\beta}_{1k} + 2\mathbf{B}_k\mathbf{x}\right] + 2\lambda(\mathbb{B})\mathbf{x} \end{array} \right\} = \mathbf{0}$$

$$\nabla_\lambda L(\mathbf{x}, \lambda; \boldsymbol{\beta}) = ||\mathbf{x}||^2 - c^2 = 0$$

□

# Conclusions

As the reader can check, the results of the paper can be computed easily from the estimates of the parameters obtained through multivariate multiple regression and a known explicit form of the functional $f(\cdot)$. A few basic routines in software R or MATLAB shall be sufficient for achievement this objective.

In addition, as a consequence of Theorem 2 now is feasible to establish confidence regions and intervals and hypothesis tests on the critical point, see (Bishop *et al.*, 1991, Section 14.6.4, pp. 498–500); it is also possible to identify operating conditions as regions or intervals, instead of isolated points.

The results of this paper can be taken as a good first approximation to the exact problem. However, in some applications the number of observations can be relatively small and perhaps the results obtained in this work should be applied with caution.

# Acknowledgements

The authors wish to thank the Editor and the anonymous reviewers for their constructive comments on the preliminary version of this paper. This paper was written during J. A. Díaz-García's stay as a professor at the Department of Statistics and O. R. of the University of Granada, España. F. J. Caro-Lopera was supported by project No. 158 of University of Medellin.

# References

[1] Aitchison, J. and S. D. Silvey, S. D. (1958): Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, **29**, 813–828.

[2] Biles, W. E. (1975): A response surface method for experimental optimization of multi-response process. *Industrial & Engeneering Chemistry Process Design Development*, **14**, 152-158.

[3] Gigelow, J. H. and Shapiro, N. Z. (1974): Implicit function theorem for mathematical programming and for systems of iniqualities. *Mathematical Programming*, **6(2)**, 141–156.

[4] Bishop, Y. M. M., Finberg, S. E. and Holland, P. W. (1991): *Discrete Multivariate Analysis: Theory and Practice*. The MIT press, Cambridge.

[5] Chatterjee, S. and Hadi, A. S. (1988): *Sensitivity Analysis in Linear Regression*. John Wiley: New York.

[6] Cramér, H. (1946): *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[7] Díaz García, J. A. and Ramos-Quiroga, R. (2001): An approach to optimization in response surfaces. *Communication in Statatistics, Part A- Theory and Methods*, **30**, 827–835.

[8] Díaz García, J. A. and Ramos-Quiroga, R. (2002): Erratum. An approach to optimization in response surfaces. *Communication in Statatistics, Part A- Theory and Methods*, **31**, 161.

[9] Dupačová, J. (1984): Stability in stochastic programming with recourse-estimated parameters. *Mathematical Programming*, **28**, 72–83.

[10] Fiacco, A. V. and Ghaemi, A. (1982): Sensitivity analysis of a nonlinear structural design problem. *Computers & Operations Research*, **9(1)**, 29–55.

[11] Jagannathan, R. (1977): Minimax procedure for a class of linear programs under uncertainty. *Operations Research*, **25**, 173–177.

[12] Kazemzadeh, R. B., Bashiri, M., Atkinson, A. C. and Noorossana, R. (2008): A General Framework for Multiresponse Optimization Problems Based on Goal Programming. *European Journal of Operational Research*, **189**, 421-429.

[13] Khuri, A. I. and Conlon, M. (1981): Simultaneous optimization of multiple responses represented by polynomial regression functions. *Technometrics*, **23**, 363–375.

[14] Khuri, A. I. and Cornell, J. A. (1987): *Response Surfaces: Designs and Analysis*. Marcel Dekker, Inc., NewYork.

[15] Miettinen, K. M. (1999): *Non linear multiobjective optimization*. Kluwer Academic Publishers, Boston.

[16] Muirhead, R. J. (1982): *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., 1982.

[17] Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2009): *Response surface methodology: process and product optimization using designed experiments*. Third edition, Wiley, New York, .

[18] Rao, C. R. (1973): *Linear Statistical Inference and its Applications*. (2nd ed.) John Wiley & Sons, New York.

[19] Rao, S. S. (1979): *Optimization Theory and Applications*. Wiley Eastern Limited, New Delhi.

[20] Ríos, S., Ríos Insua, S. and Ríos Insua, M. J. (1989): *Procesos de decisión Multicriterio*. EUDEMA, Madrid, (in Spanish).

[21] Steuer, R. E. (1986): *Multiple criteria optimization: Theory, computation and applications*. John Wiley, New York.

Metodološki zvezki, Vol. 12, No. 1, 2015, 25-43

# Empirical Option Weights for Multiple-Choice Items: Interactions with Item Properties and Testing Design

Gregor Sočan[1]

**Abstract**

In scoring of a multiple-choice test, the number of correct answers does not use all information available from item responses. Scoring such tests by applying empirically determined weights to the chosen options should provide more information on examinees' knowledge and consequently produce more valid test scores. However, existing empirical evidence on this topic does not clearly support option weighting. To overcome the limitations of the previous studies, we performed a simulation study where we manipulated the instruction to examinees, discrimination structure of distractors, test length, and sample size. We compared validity and internal consistency of number-correct scores, corrected-for-guessing scores, two variants of correlation-weighted scores and homogeneity analysis scores. The results suggest that in certain conditions the correlation-weighted scores are notably more valid than the number-correct scores. On the other hand, homogeneity analysis cannot be recommended as a scoring method. The relative performance of scoring methods strongly depends on the instructions and on distractors' properties, and only to a lesser extent on sample size and test length.

## 1 Introduction

The multiple-choice format is a popular item format for ability and attainment tests, especially when a maximally objective or even an automated scoring is desired, or when the use of constructed-response items would be impractical. However, the relatively complex form of a multiple-choice item allows for

[1] Gregor Sočan, Department of psychology, University of Ljubljana, Aškerčeva c. 2, SI-1000 Ljubljana, Slovenia; gregor.socan@ff.uni-lj.si

different scoring techniques, which differ with regard to the information they take into account and the specific algorithm for transforming this information into a single numeric score. Since guessing can be an important factor contributing to the test score, the instructions for examinees and the scoring algorithm should not be sensitive to individual differences in the guessing-related attitudes and processes. This paper deals with the practical usefulness of scoring techniques, based on empirically derived weights, which are consistent with the classical test theory (thus excluding item response theory models). Furthermore, we shall limit our treatment to the most common form of items, where the examinee is required to choose one of the $m$ offered alternatives, $m > 2$, excluding item formats like true/false, answer-until-correct, option elimination, confidence weighting etc.

## 1.1 Scoring based on the number of correct responses only

The simplest scoring rule is the number-correct (NC) rule: the score is determined as a number of correct responses. The most obvious advantage of the NC rule is its conceptual and computational simplicity. On the other hand, the test scores are contaminated with the examinees' tendency to guess, unless all examinees respond to all items. Moreover, the NC rule does not use all information available from the item response. In particular, when the examinee does not respond correctly, (s)he gets zero points regardless of the degree of incorrectness of the chosen distractor.

To eliminate the effect of individual differences in the guessing proneness, the correction for guessing (CG), also known as "formula scoring" can be used. This rule assigns different scores to incorrect responses and to omitted responses; typically, an incorrect response is scored with $-[1/(m-1)]$ points, and an omitted response is scored with 0 points[2]. Although the CG scoring could be treated as a simple case of differential response weighting, this conceptualization would imply that the differences between 1, 0 and $-[1/(m-1)]$ points reflect the differences in expected values of the respective examinees' knowledge – a position that would be disputable at best. It is therefore more meaningful to treat it as a modification of the NC scoring. In any case, the opinions about the correction for guessing have varied widely (for a recent review see Lesage, Valcke, and Sabbe, 2013). In short, the advocates have mostly stressed the higher measurement precision relative to the NC scores (Burton, 2004, 2005; Lord, 1975), which is achieved when some basic assumptions about the response process hold. On the other hand, the critics (for instance, Bar-Hillel, Budescu, and Attali, 2005; Budescu and Bo, 2015; Cross and Frary, 1977) pointed to introduction of biases related to the individual differences in risk aversion, accuracy of the subjective estimation of the probability of a correct response, and similar psychological factors. The issue of

[2] Alternative CG formulas and algorithms have been proposed; for an evaluation see Espinosa and Gardeazabal (2010, 2013).

correction for guessing is complicated by the finding that statistically equivalent variants of the CG scoring are not necessarily strategically equivalent when the risk aversion is taken into account (Espinosa and Gardeazabal, 2013), and by the fact that researchers working with similar quantitative models of examinees' responding to multiple choice items sometimes arrived to opposite conclusions about the optimal size of the penalty for a wrong answer (Budescu and Bo, 2015 vs. Espinosa and Gardeazabal, 2013). We shall not extend our discussion of the CG scoring, because it does not play the central role in this study; however, we should stress that various irrelevant psychological factors may significantly determine the psychometric properties of guessing-corrected scores.

## 1.2 Empirical option weighting

The previously discussed scoring rules disregard the information contained in the particular choice of an incorrect option. That is, if the examinee chooses an incorrect option, the item score does not depend on which option has been chosen. If the distractors differ in the level of incorrectness, which may often be the case, such scoring does not use all information contained in the item response. Therefore, taking account of a choice of a particular distractor should in principle increase reliability and validity of the examinee's score. One possible way of using this information is by using an item-response theory (IRT) model, modelling the relation between the latent trait and the response to each option. Such models include cases described by Bock (1997), Thissen and Steinberg (1997), Revuelta (2005), and García-Pérez and Frary (1991). However, if the sample size is not very large or when the assumptions of the IRT models are not satisfied – for instance, in areas like educational measurement or personality assessment the measured construct is often multidimensional – researchers may prefer to use classical models, based on linear combinations. A computational approach which seems particularly attractive is homogeneity analysis (HA; also known as dual scaling, optimal scaling, multiple correspondence analysis, and Guttman (1941) weighting). Homogeneity analysis transforms qualitative variables into quantitative variables by means of weighting the elements of the indicator matrix, corresponding to the item options.

In case of multiple choice items, the responses to each of $k$ items are recorded as a categorical variable with $m$ categories (or possibly $m + 1$ categories, if omissions are to be weighted as well). This variable is then recoded into $m$ (or $m + 1$, respectively) indicator variables, taking the value of 1 if the corresponding option was selected by the examinee, and the value of 0 otherwise. Then, $k$ vectors of weights are computed minimizing the discrepancy function

$$f(\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{h}) = \sum_{j=1}^{k} \left\| \mathbf{G}_j \mathbf{w}_j - \mathbf{h} \right\|^2 \quad (1)$$

where $\mathbf{G}_j$ is a matrix of indicator variables for item *j*, and **h** is a vector of weighted test scores, whereas the symbol $\|\mathbf{X}\|^2$ stands for the sum of squared elements of the matrix **X**. Since Equation 1 formally represents a multivariate regression model, we can interpret the score vector **h** as the best linear approximation of the quantified item responses in the least squares sense. That is, the score explains the maximum possible amount of quantified items' variances. Conversely, the quantified variables are maximally homogeneous, which means that they have minimum within-person variance. Although several uncorrelated score vectors can be determined, only the first one is interesting in our case.

From the psychometric viewpoint, a particularly attractive feature of HA is that its weights maximize the value of coefficient alpha for the total test score, calculated as the sum of the quantified variables. From the interpretational viewpoint, HA is attractive because of its close relation to the principal component analysis (PCA). HA can be understood as a variation of PCA for categorical variables; note that the first principal component also maximizes the explained variances of the analyzed variables and the coefficient alpha for their sum (for further details see Gifi, 1990, Greenacre, 2007, and Lord, 1958).

Two potential drawbacks of HA should be noted. As a statistical optimization technique, it is prone to chance capitalization, and it is not clear what sample sizes are needed for a reliable generalization of multiple-choice item weights. Furthermore, although it makes intuitive sense to expect a high correlation between the score vector and the measured knowledge, there is no guarantee for this to happen in a particular dataset. Fortunately, the seriousness of both potential problems can be assessed by means of simulation.

A compromise solution, sacrificing optimality to achieve greater stability, would consist of weighting the indicator variables with their point-biserial correlations with the total test score. The correlation weights (CW) should generally result in a smaller amount of quantified items' homogeneity, but on the other hand their use should reduce the scope of the above mentioned limitations of HA. Since the correlation weights are not aimed at maximizing homogeneity in the sample, they do not capitalize on chance; and because the total test score is always relatively highly correlated with knowledge, the weighted score cannot have a very low correlation with knowledge. Although not producing optimal scores, the correlation weights are appealing because they still give more importance to better discriminating options. Furthermore, the calculation of correlations with the total unweighted score represents the first step in Guttman's (1941) version of HA. Although this computational procedure for HA is now obsolete, this fact shows that CW can be understood as the first-step approximation to the HA solution – similarly as the classical item discrimination coefficients are the first-step approximations to the first principal component loadings.

Both approaches to the empirical option weighting were evaluated in several older studies, which mostly focused on the internal-consistency reliability and the

criterion validity of the weighted option scores in large samples. Davis and Fifer (1959) found that the cross-validated correlation weights, compared to the NC scores, increased reliability, but not validity. Reilly and Jackson (1973), Downey (1979), and Hendrickson (1971) reported that the HA scores, compared to the NC scores, were more reliable and less valid. Sabers and Gordon (1969) found no notable differences in either reliability or validity. Echternacht (1976), on the other hand, reported higher values of both reliability and validity coefficients for the HA scores compared to the NC scores.

Almost all studies found higher internal consistency of HA scores compared to NC scores; however, this should be expected due to the alpha-maximizing property of HA. Inconsistent results with regard to criterion validity are more puzzling. Echternacht (1976) simply attributed the inconsistency to the fact that different criterion variables implied different validity coefficients. Hendrickson (1971) conjectured that lower correlations with other variables were a consequence of a more homogeneous factor structure of HA items scores relative to the unweighted item scores.

Reilly and Jackson (1973) noted that the weights assigned to omitted responses were often very low in practice. Raffeld (1975) assigned a constant zero weight to omitted responses, which improved predictive validity of the weighted option scores. However, this solution does not seem completely satisfactory, because setting the value of a weight to zero does not have a clear rationale. Nevertheless, Raffeld's results are important because they highlight the issue of weighting the omitted responses.

The inconsistent results concerning validity of option-weighted scores eventually led Frary (1989) to conclude that "option weighting offers no consistent basis for improving the psychometric quality of test scores unless there is a problem with respect to internal consistency reliability." (p. 83). However, the preceding empirical studies had a common limitation: for validity assessment, they relied on the correlation with external criterion variable(s). A more definite conclusion could be reached if the correlation with the response-generating trait was determined. Sočan (2009) simulated test scores while manipulating the degree of incorrectness of distractors and the general level of guessing. The results predictably showed better homogeneity and reliability of both CW and HA scores, compared to the NC scores, while the validity (i.e., correlation with the latent trait) of the CW scores was either marginally worse or notably better than the validity of the NC scores, the difference being larger when there was a distractor with a positive discrimination (for instance, a partially correct alternative). However, Sočan's study had some limitations. First, only performance in small samples was studied, and the weights were neither cross-validated nor generalized to the population. Second, risk-aversion / guessing-proneness was not included as a person characteristic, which is not consistent with empirical evidence indicating the existence of guessing tendency as a stable personality-related trait (for instance, Brenk and Bucik, 1994; Dahlbäck, 1990; Slakter, 1967).

The problem of this study was to assess the psychometric quality (especially construct validity) of the correlation-weighted (CW) and homogeneity analysis (HA) scores compared to the number-correct scores (NC) and "guessing-corrected" (CG) scores. For all types of weighted option scores, we aimed to determine the generalizability of sample weights to the population with relation to sample size, test length, and distractor characteristics.

# 2 Method

## 2.1 Experimental treatments

The following four experimental conditions were manipulated in a crossed design.
1. Test instructions. In the first level of the factor ("forced choice"), simulated examinees were forced to choose one of the alternatives, regardless of their confidence in the correctness of their choice. In the second level ("omissions allowed"), they were instructed to respond only when being reasonably confident that the selected alternative was correct.
2. Distractor properties. Each test consisted of multiple-choice items with three alternatives. This format was chosen both for simplicity and according to recommendations from the meta-analysis by Rodriguez (2005). The alternatives were characterized by its discriminating power, defined as the correlation with the measured trait (denoted by $r_{j(a)\theta}$). The discrimination parameter for the correct option was always $r_{j(c)\theta}$ = .40. The discrimination parameters for the two distractors were different in the two levels of the factor, namely:
   1. "both negative": $r_{j(1)\theta}$ = -.20 and $r_{j(2)\theta}$ = -.10;
   2. "one positive": $r_{j(1)\theta}$ = -.20 and $r_{j(2)\theta}$ = .10.

The first factor level corresponds to the case where both distractors are clearly incorrect, while the second factor level corresponds to the case with one partially correct distractor. In both cases, all items in the test had equal discrimination structure. The values of parameters were set according to the results of preliminary analyses on real data (not shown here).
3. Test length. Two test lengths were used, with the number of items $k$ = 15 and 50, respectively.
4. Sample size. Three sample size levels were used: $n$ = 100, 200 and 500, respectively.

## 2.2 Simulation design

In the first step, a population was defined as a group of one million examinees, characterized by two uncorrelated latent traits: knowledge ($\theta \sim N(0,1)$) and risk-aversion ($\gamma \sim N(0,0.5)$). The variance of risk-aversion was smaller than the variance of knowledge to prevent the presence of examinees with extreme guessing/omitting behavior (see the description of the response model below). Responses were generated for each of the eight combinations of factors 1-3.

We used the underlying variables approach to the response generation. For each examinee, we computed his/her response propensity for each option:

$$x_{ij(a)}^{*} = \theta_i r_{j(a)\theta} + e_{ij(a)}, \quad (2)$$

where $e_{ij(a)}$ was a random error component, $e_{ij(a)} \sim N(0,(1-r^2_{j(a)\theta})^{1/2})$.

In the "forced choice" condition, the response was simply determined as the option with the largest response propensity. In the "omissions allowed" condition, an examinee responded only when $\max(x_{ij(a)}^{*}) > \gamma_i$, and omitted the response otherwise. Therefore, our response model assumes that an examinee finds the option which seems to be correct most likely, but omits the response (if allowed to do so) if his/her confidence in the correctness of the choice does not exceed the personal confidence criterion (which is higher for more risk-averse persons and vice versa).

When all responses were generated, 10000 random samples of each size were drawn without replacement from each of the population response matrices. Altogether, $3 \times 8 \times 10000 = 240000$ samples were processed. In each sample, both correlation weights and homogeneity weights were determined. Two versions of correlation weights were calculated. In both cases the weights were calculated as Pearson correlations between the indicator variable corresponding to an option (including, where relevant, an omission) and the test score. The first version ($CW_{NC}$) used the number-correct score as the criterion variable, and the second version ($CW_{CG}$) used the score corrected for guessing. The HA weights were calculated using the closed form solution for homogeneity analysis (ten Berge, 1993, pp. 66-67). For the CG scoring, the standard penalty of $-[1/(m-1)] = -0.5$ points for an incorrect response was used. Five sets of scores were thus obtained for each sample, and the sample values of criterion variables were calculated. Then, the CW and HA weights obtained in the sample were applied to the population responses, and the values of criterion variables were calculated again. Table A in the appendix presents percentages of the choices and both validities and coefficients alpha for the number-correct score in various conditions.

All simulations were performed using MATLAB R2012b software. The MATLAB codes used for the simulations are available as supplementary material from the website of Metodološki zvezki (http://www.stat-d.si/mz/Articles.html).

# 3 Results

We shall first compare the four scoring rules with respect to validity and internal consistency of the obtained scores. After that, we shall compare both instruction conditions with respect to the validity of scores. We present no inferential statistics: because of large sample sizes and due to the use of the repeated-measures design, the statistical power was very high, and even some practically irrelevant effects reached statistical significance. All discussed results were statistically significant ($p < .001$).

## 3.1 Validity

The most important question addressed in this study is validity of scores obtained by different scoring rules. Since the responses were simulated, it was possible to assess the construct validity directly as the Pearson correlation between the test score and the latent knowledge. Table 1 presents validity increments or losses, respectively, i.e. differences between mean validity coefficients of the number-correct score and mean validity coefficients of each of the three remaining scores in various conditions. A positive difference indicates that a scoring rule produces more valid scores than the number-correct rule. The values in the left part of the table are related to the sample validities, while the values in the right part of the table are computed from the generalized validities; that is, they are based on population scores computed with sample weights. Since the CG weights are not empirically estimated, the respective population validities are independent of sample size.

Clearly, the instruction for examinees was a major determinant of the pattern of validity increments. When examinees could omit the response, both the CG scores and the $CW_{CG}$ scores were invariably more valid than the NC scores. As expected, when there was a positively discriminating distractor, option weighting resulted in a slightly higher increment than the simple correction for guessing. On the other hand, when both distractors had negative discrimination power, the guessing correction produced higher increments than option weights, unless the sample size was at least 500. For both scoring rules, the increments were higher for longer tests. The remaining two weighting schemes produced scores less valid than the number-correct scores. In case of the HA scoring, the differences were especially large; in fact, the average validity of the HA scores was lower than .40 in all 12 conditions related to the "omissions allowed" instruction. The differences for the $CW_{NC}$ scores were much smaller, but they were still of notable size and consistently negative. The increments and losses, respectively, of all four scoring methods were larger in absolute value when the test was longer.

In the forced choice condition, the increments for the CG scoring are not presented, because the CG scores are linearly related to the NC scores and are therefore equally valid. Consequently, the $CW_{CG}$ rule and the $CW_{NC}$ rule are also equivalent.

**Table 1:** Validity increments/losses relative to the NC scoring

| | | | | Sample | | | | Population | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ins. | *k* | $r_{j\theta} > 0$ | *n* | CG | $CW_{NC}$ | $CW_{CG}$ | HA | CG | $CW_{NC}$ | $CW_{CG}$ | HA |
| FC | 15 | one | 100 | - | **.02** | -- | -.02 | - | **.02** | -- | -.01 |
| | | | 200 | - | **.03** | -- | **.02** | - | **.03** | -- | **.02** |
| | | | 500 | - | **.03** | -- | **.04** | - | **.03** | -- | **.04** |
| | | none | 100 | - | -.01 | -- | -.04 | - | -.01 | -- | -.04 |
| | | | 200 | - | .00 | -- | -.02 | - | .00 | -- | -.02 |
| | | | 500 | - | **.00** | -- | .00 | - | **.00** | -- | .00 |
| | 50 | one | 100 | - | **.01** | -- | **.01** | - | **.01** | -- | **.01** |
| | | | 200 | - | **.02** | -- | **.02** | - | **.02** | -- | **.02** |
| | | | 500 | - | **.02** | -- | **.02** | - | **.02** | -- | **.02** |
| | | none | 100 | - | -.01 | -- | -.01 | - | -.01 | -- | -.01 |
| | | | 200 | - | .00 | -- | .00 | - | .00 | -- | .00 |
| | | | 500 | - | **.00** | -- | .00 | - | **.00** | -- | .00 |
| OA | 15 | one | 100 | **.02** | -.02 | **.04** | -.42 | ↑ | -.02 | **.03** | -.44 |
| | | | 200 | **.02** | -.01 | **.05** | -.43 | **.02** | -.02 | **.05** | -.43 |
| | | | 500 | **.02** | -.01 | **.05** | -.43 | ↓ | -.01 | **.05** | -.43 |
| | | none | 100 | **.04** | -.03 | **.02** | -.52 | ↑ | -.03 | **.02** | -.58 |
| | | | 200 | **.04** | -.02 | **.03** | -.57 | **.04** | -.03 | **.03** | -.60 |
| | | | 500 | **.04** | -.02 | **.04** | -.59 | ↓ | -.02 | **.04** | -.60 |
| | 50 | one | 100 | **.05** | -.06 | **.06** | -.52 | ↑ | -.08 | **.05** | -.54 |
| | | | 200 | **.05** | -.07 | **.06** | -.53 | **.05** | -.08 | **.06** | -.53 |
| | | | 500 | **.05** | -.07 | **.06** | -.53 | ↓ | -.07 | **.06** | -.53 |
| | | none | 100 | **.05** | -.05 | **.04** | -.64 | ↑ | -.06 | **.04** | -.69 |
| | | | 200 | **.05** | -.05 | **.05** | -.67 | **.05** | -.05 | **.04** | -.70 |
| | | | 500 | **.05** | -.05 | **.05** | -.69 | ↓ | -.05 | **.05** | -.69 |

Positive values are in boldface. Ins. = instruction, FC = forced choice, OA = omissions allowed, *k* = number of test items, *n* = sample size, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, CG = correction for guessing, $CW_{NC}$ = correlation weights (number-correct score as the criterion), $CW_{CG}$ = correlation weights (guessing-corrected score as the criterion), HA = homogeneity analysis weights, - not relevant because both rules are equivalent, -- not reported because it is equal to the $CW_{NC}$ value, ↑ equal to the value below, ↓ equal to the value above.

With the forced choice instruction, the average validities of the weighted option scores were generally higher than the validities of the NC scores when one distractor had a positive discrimination. For the CW scores, the validity increment was slightly higher for the shorter test than for the longer test. When both

distractors discriminated negatively, the differences were close to zero, except when the sample size was 100; in this case, the weighted option scores were less valid than the NC scores, and we may conjecture that the weighted option scores would be more valid than the NC scores in larger samples (i.e., larger than 500). In general, validity increments of the weighted option scores were larger in larger samples, however, the effect of the sample size was slim.

We can also note two general observations. First, the pattern of the sample differences was very similar to the pattern of the population differences. Therefore, for the evaluation of validity of scores obtained by various scoring methods, it does not matter much whether only a particular sample is of interest, or the generalization to the population is desired. Second, the average validity of (both variants of) CW scores was at least as high as the average validity of HA scores in almost all conditions, and it was much higher when it was possible to omit the response.

## 3.2 Internal consistency

In the broad sense, internal consistency is related to the homogeneity of a group of measures, in this case test items. We evaluated two aspects: internal consistency reliability and the amount of items' variance, explained by the first principal component. For the former, we used coefficient alpha, which should be an accurate estimate of reliability, since the items were constructed as parallel measurements. The values of coefficient alpha were calculated from the covariance matrices of the item scores. For the latter, we performed principal component analysis on the inter-item correlation matrix. We do not report sample results; it is a mathematical necessity that the HA scores are more internally consistent than the NC scores, and since the correlation weights are approximations to the HA weights, the same can be expected for both variants of CW scores. Indeed, all weighted option scores had higher sample values of both coefficient alpha and the explained variance than the NC scores.

However, when weights are cross-validated or generalized to the population, respectively, the superiority of weighted option scores is not guaranteed any longer. Table 2 presents internal consistency indicators for the scores, obtained by applying sample weights to the population response matrix. As in the previous section, the increments/losses relative to the NC scoring are presented. First, we can note a general – and, of course, expected - pattern of increasing internal consistency increments with sample size – that is, when the weights were obtained in larger samples, they generalized to the population better.

In all "omissions allowed" conditions, internal consistency of the HA scores was notably better, and internal consistency of the CG scores was somewhat worse in comparison to the NC scores. The behavior of CW scores was more complex. The $CW_{NC}$ scoring resulted in positive increments where one distractor had a

positive discrimination; otherwise, the increment was positive only when the sample size was relatively large. On the other hand, the increments for the $CW_{CG}$ scores were positive only when the test was long, the sample size was 500 and there was a positively discriminating distractor.

**Table 2:** Internal consistency increments/losses relative to the NC scoring

| | | | | α | | | | $Var_{PC1}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ins. | *k* | $r_{j\theta} > 0$ | *n* | CG | $CW_{NC}$ | $CW_{CG}$ | HA | CG | $CW_{NC}$ | $CW_{CG}$ | HA |
| FC | 15 | one | 100 | - | **.03** | -- | -.03 | - | **1.03** | -- | **0.23** |
| | | | 200 | - | **.04** | -- | **.03** | - | **1.28** | -- | **1.12** |
| | | | 500 | - | **.05** | -- | **.05** | - | **1.42** | -- | **1.64** |
| | | none | 100 | - | -.02 | -- | -.07 | - | -0.34 | -- | -1.40 |
| | | | 200 | - | -.01 | -- | -.03 | - | -0.07 | -- | -0.54 |
| | | | 500 | - | .00 | -- | -.01 | - | **0.08** | -- | -0.08 |
| | 50 | one | 100 | - | **.02** | -- | **.01** | - | **1.20** | -- | **0.90** |
| | | | 200 | - | **.03** | -- | **.03** | - | **1.59** | -- | **1.48** |
| | | | 500 | - | **.04** | -- | **.04** | - | **1.82** | -- | **1.83** |
| | | none | 100 | - | -.02 | -- | -.02 | - | -0.53 | -- | -0.88 |
| | | | 200 | - | -.01 | -- | -.01 | - | -0.15 | -- | -0.31 |
| | | | 500 | - | .00 | -- | .00 | - | **0.07** | -- | **0.01** |
| OA | 15 | one | 100 | -.07 | **.05** | -.03 | **.15** | ↑ | **2.13** | -.57 | **7.33** |
| | | | 200 | -.07 | **.06** | -.01 | **.17** | -1.82 | **2.54** | -.22 | **8.51** |
| | | | 500 | -.07 | **.07** | .00 | **.18** | ↓ | **2.78** | -.02 | **9.15** |
| | | none | 100 | -.03 | -.01 | -.05 | **.07** | ↑ | -.19 | -1.55 | **4.12** |
| | | | 200 | -.03 | .00 | -.04 | **.09** | -1.13 | **.20** | -1.18 | **5.41** |
| | | | 500 | -.03 | **.01** | -.03 | **.11** | ↓ | **.42** | -.97 | **6.10** |
| | 50 | one | 100 | -.04 | **.03** | -.02 | **.08** | ↑ | **3.04** | -.25 | **8.27** |
| | | | 200 | -.04 | **.04** | .00 | **.09** | -1.92 | **3.68** | **.30** | **9.18** |
| | | | 500 | -.04 | **.05** | **.01** | **.09** | ↓ | **4.07** | **.61** | **9.70** |
| | | none | 100 | -.02 | -.01 | -.03 | **.04** | ↑ | -.16 | -1.82 | **5.01** |
| | | | 200 | -.02 | **.00** | -.02 | **.05** | -1.18 | **.41** | -1.30 | **5.99** |
| | | | 500 | -.02 | **.01** | -.01 | **.05** | ↓ | **.75** | -.98 | **6.53** |

Positive values are in boldface. α = coefficient alpha of internal consistency reliability, $Var_{PC1}$ = percentage of variance explained by the first principal component, *k* = number of test items, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, *n* = sample size, CG = correction for guessing, $CW_{NC}$ = correlation weights (number-correct score as criterion), $CW_{CG}$ = correlation weights (guessing-corrected score as criterion), HA = homogeneity analysis weights, - not relevant because both rules are equivalent, ↑ equal to the value below, ↓ equal to the value above.

In the "forced choice" conditions, the weighted option scores were in general more internally consistent in cases with a positively discriminating distractor; otherwise, the internal consistency of weighted option scores was comparable to the internal consistency of the NC scores when the weights were obtained in

samples of size 500, and somewhat smaller when obtained in smaller samples. Therefore, when examinees are required to choose an option, using empirical option weights might significantly increase the population internal consistency only in samples of size considerably larger than 500. The performance of the HA scoring was generally comparable or even slightly worse than the performance of the CW scoring.

## 3.3 Effect of instructions on validity

**Table 3:** Validity increments of the "forced choice" over the "omissions allowed" instructions

| | | | Sample | | | | | Population | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $r_{j\theta} > 0$ | $n$ | NC | CG | $CW_{NC}$ | $CW_{CG}$ | HA | NC | CG | $CW_{NC}$ | $CW_{CG}$ | HA |
| 15 | one | 100 | .03 | .01 | .07 | .01 | .43 | ↑ | ↑ | .07 | .01 | .46 |
| | | 200 | .03 | .01 | .07 | .01 | .48 | .03 | .01 | .07 | .01 | .48 |
| | | 500 | .03 | .01 | .07 | .01 | .49 | ↓ | ↓ | .07 | .01 | .50 |
| | none | 100 | .05 | .01 | .07 | .01 | .53 | ↑ | ↑ | .07 | .02 | .58 |
| | | 200 | .05 | .01 | .07 | .01 | .60 | .05 | .01 | .07 | .01 | .63 |
| | | 500 | .05 | .01 | .07 | .01 | .63 | ↓ | ↓ | .07 | .01 | .64 |
| 50 | one | 100 | .06 | .01 | .14 | .02 | .59 | ↑ | ↑ | .16 | .03 | .61 |
| | | 200 | .06 | .01 | .15 | .02 | .60 | .06 | .01 | .16 | .02 | .61 |
| | | 500 | .06 | .01 | .15 | .02 | .61 | ↓ | ↓ | .16 | .02 | .61 |
| | none | 100 | .06 | .01 | .11 | .01 | .69 | ↑ | ↑ | .11 | .02 | .74 |
| | | 200 | .06 | .01 | .11 | .01 | .73 | .06 | .01 | .11 | .02 | .76 |
| | | 500 | .06 | .01 | .11 | .01 | .75 | ↓ | ↓ | .11 | .02 | .76 |

$k$ = number of test items, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, $n$ = sample size, NC = number-correct, CG = correction for guessing, $CW_{NC}$ = correlation weights (number-correct score as criterion), $CW_{CG}$ = correlation weights (guessing-corrected score as criterion), HA = homogeneity analysis weights, ↑ equal to the value below, ↓ equal to the value above.

A question can be posed whether one type of instruction generally results in a higher average validity, and whether this effect is moderated by the choice of a scoring method. Table 3 presents the average differences between validity coefficients in the "forced choice" condition and the same coefficients in the "omissions allowed" condition. The presented results show that the instructions for examinees strongly determine the performance of various scoring methods. All differences were positive: the average validity in a specified test design was always higher in the "forced choice" than in "omissions allowed" condition. The sizes of the average differences were very similar for both sample and population validities. The differences were in general larger for longer tests, whilst the effect of sample size was negligible. Although the instruction effect was present in all

types of scores, the average size of the difference varied: it was quite small for the CG and $CW_{CG}$ scores, and very large for the HA scores; in the latter case, the extreme superiority of the "forced choice" instruction was related to the very poor validity of the HA scores in the "omissions allowed" condition, as reported in section 3.1.

# 4 Discussion

We see the main contribution of this study in elucidating the interactions between various aspects of testing design and the performance of different scoring methods. If examinees are instructed to attempt every item, the expected validity is higher than if they are instructed to avoid guessing. This fact reveals that individual differences in guessing behavior are generally more detrimental to measurement quality than the guessing itself. Although the instruction effect is present generally, its size depends on the scoring method and - to a smaller extent – on the test length. The effect of test length can be attributed to the law of large numbers: with a larger number of items, the proportion of "lucky guesses" converges to its expectation. Lengthening the test therefore reduces the effect of luck, but does not affect the effect of personality factors of guessing behavior.

When examinees answer all items (as in the "forced-choice" condition), the number-correct scoring should be preferred to the option-weighting scoring if all distractors have negative discriminations of similar size; option-weighted scores seem to be more valid only in samples much larger than 500. On the other hand, when partially correct distractors (with a small positive discrimination power) are included, option weighting increases validity compared to the number-correct scoring, even in small samples (like $n$=100).

When guessing is discouraged and examinees omit some items (the "omissions allowed" condition), the correction for guessing and correlation weighting (based on the guessing-corrected scores) should be preferred methods. The $CW_{CG}$ scoring may be the method of choice if there is a positively discriminating distractor; when all distractors have negative discrimination, the CG scores are more valid than the $CW_{CG}$ scores if the sample size is less than about 500. Note than the $CW_{CG}$ rather $CW_{NC}$ weights should be used with this instruction. Because the CW scoring uses more information than the CG scoring, we speculate that it may be less sensitive to personality factors like subjective utility (as discussed in Budescu and Bo, 2015, and Espinosa and Gardeazabal, 2013). When both types of scores have similar validity and reliability, it may be thus safer to use the $CW_{CG}$ scores.

According to our results, homogeneity analysis cannot be recommended as a scoring method. The HA scores were never notably more valid than the correlation-weighted scores, and they were substantially less valid when examinees could omit items. Obviously, the relatively high internal consistency of

the HA scores in the "omissions allowed" condition does not imply a high validity, but only reflects the contamination of true scores with the non-cognitive variables determining the guessing level. In this condition, the classical true score is not a faithful representation of knowledge, but rather a conglomerate of knowledge and risk-aversion. Therefore, the HA score may be a highly reliable measure of a conceptually and practically irrelevant quantity. Clearly, reliability is not a relevant test characteristic in such circumstances. As a collateral contribution, our results thus illustrate the danger of relying solely on (internal consistency) reliability in evaluation of measurement quality of cognitive tests.

The unsatisfactory behaviour of the HA weights may be surprising. It can be partly attributed to the fact that the HA algorithm does not assure a high correlation between the weighted score and the number correct score, which is especially problematic when there are other significant determinants of the examinee's response (in particular, the guessing proneness). The performance of the HA weights was better when the guessing proneness was controlled (by forcing all examinees to answer all items); although their validity was comparable to the validity of their approximations (i.e., the CW weights) in our sample size range, we may speculate that the HA weights might be superior in terms of validity in very large samples (for instance, in large standardized educational assessments), provided that both empirical scoring key and the forced-choice instruction would be considered acceptable.

In the "omissions allowed" conditions, the $CW_{NC}$ and HA weights on one hand and the $CW_{CG}$ weights on the other hand performed markedly different with respect to validity. These differences should be attributed to the differences in treatment of omitted responses. The number-correct score, controlled for the level of knowledge, is higher for examinees with a lower level of risk-aversion (that is, for examinees who attempt more items). As a consequence, a $CW_{NC}$ weight for an omitted response is negative even if the actual decision between responding and omitting is completely unrelated to the level of knowledge. Indeed, the inspection of sample weights for omissions (not reported here) showed that the values of $CW_{NC}$ weights were typically close to the weights for the (negatively discriminating) distractor(s), while the corresponding $CW_{CG}$ weights were much closer to zero (cf. Reilly and Jackson, 1973). With the homogeneity analysis weights, the problem is essentially the same, but is aggravated due to a lack of a mechanism that would ensure at least approximate collinearity with knowledge. Using $CW_{CG}$ weights removes the negative bias from the omission weights, however, these weights still reflect only one determinant of a response omission (i.e., knowledge), and disregard the personality-related determinants (risk aversion, subjective utility of a score, self-confidence and so on).

Increasing sample size seems to improve the performance of correlation-weighted scores. However, the sample size effect was quite small overall: weights determined in samples of 100 persons did not generalize substantially less well

than the weights determined in samples of 500 persons, especially when the test was long ($k$=50).

It should be noted that the patterns rather than particular values of the reported effects are to be taken as findings of this study. The values of the increments/losses depend on particular experimental settings, which are to a certain extent arbitrary. For instance, we can reasonably expect that using much easier items would make option-weighting less relevant because of a smaller percentage of incorrect responses. On the other hand, distractors with more heterogeneous discriminations would probably make the performance of the CW scores more favorable relative to the NC scores. Increasing sample size also seems to improve the performance of the CW scoring. However, a researcher who has collected a very large sample may first try to fit a more sophisticated item-response theory scoring model.

The research of scoring models for multiple-choice items is complicated by the lack of formalized cognitive models explaining the item response process. The existing models (for instance, Budescu and Bo, 2015, Espinosa and Gardeazabal, 2013) are mainly focused on guessing behavior, and are not useful in the context of weighting item options. Our study was based on very simple response model, which predicts examinee's decisions with just two person parameters (knowledge and risk-aversion). The model rests on two assumptions:

1. Both knowledge and risk-aversion are stable personal properties, which do not change during the testing process.
2. An examinee's response to a multiple-choice item is based on a comparison of plausibility of offered alternatives. When guessing is discouraged, the examinee omits the response if none of the option plausibility estimates exceeds the subjective criterion, determined by his/her level of risk-aversion.

While it seems safe to generally accept the first assumption, the second assumption is not so evidently generalizable to all examinees and testing contexts, and should be tested empirically in the future.

Empirical weighting should not be used with negatively discriminating items. Although the resulting negative weights for correct responses would generally increase both validity and internal consistency of the empirically weighted scores in comparison to the NC and CG scores, respectively, such scoring would be difficult to justify to test takers, developers and administrators. Of course, the presence of negatively discriminating items is problematic regardless of which scoring method has been chosen.

Furthermore, the weighting techniques assume invariance of item-option discrimination parameters across examinee subgroups. In real applications, this assumption may sometimes be violated. For instance, in intelligence testing, the discrimination parameters might be different for groups of examinees using different problem-solving strategies. In educational testing, using different textbooks or being exposed to different teaching methods might also cause

differences in discrimination parameters. Fortunately, the validity of the invariance assumption can be empirically verified, if the potentially critical subgroups can be identified.

# 5 Conclusions

Our results confirm Nunnally and Bernstein's (1994, p. 346) recommendation to instruct examinees to attempt every item. This instruction should not be questionable in psychological testing, especially when applying computerized tests, which can prevent possible accidental omissions. Correlation weights can be used to maximize the score validity if the distractors differ in their degree of incorrectness. In educational testing[3], some test administrators may not be comfortable with forcing the students to guess when they really do not recognize the correct answer. Using correlation weights, based on the corrected-for-guessing sum score, can be recommended in this case, especially if partially correct distractors have been used and the sample size is not too small. However, because omissions depend on both knowledge and risk-aversion, the scoring schemes studied here do not provide optimal scores for the omitted responses. Consequently, the validity of scores obtained with this instruction is lower compared to the "forced-choice" scores. Development of scoring models which would incorporate information on examinees' risk-aversion and other relevant personal characteristics, remains a task for future research.

# References

[1] Bock, R.D. (1997): The nominal categories model. In W. van der Linden and R. K. Hambleton (Eds): *Handbook of Modern Item Response Theory,* 33-49. New York: Springer.

[2] Brenk, K. and Bucik, V. (1994): Guessing of answers in objective tests, general mental ability and personality traits according to 16-PF questionnaire. *Review of Psychology*, **1**, 11-20.

[3] Budescu, D.V. and Bo, Y. (2015): Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, **80**, 1105-1122.

[3] It should be stressed that the methods discussed here are optimal in the context of the normative measurement, where the individual differences in knowledge are of main interest. Criterion-referenced measurement (for instance, a typical classroom assessment of knowledge) may call for different approaches to scoring.

[4] Burton, R.F. (2004): Multiple choice and true/false tests: Reliability measures and some implications of negative marking. *Assessment and Evaluation in Higher Education*, **29**, 585-595.

[5] Burton, R.F. (2005): Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education*, **30**, 65-72.

[6] Bar-Hillel, M., Budescu, D., and Attali, Y. (2005): Scoring and keying multiple choice tests: A case study in irrationality. *Mind and Society*, **4**, 3-12.

[7] Cross, L.H. and Frary, R.B. (1977): An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement*, **14**, 313-321.

[8] Dahlbäck, O. (1990): Personality and risk-taking. *Personality and Individual Differences*, **11**, 1235–1242.

[9] Davis, F.B. and Fifer, G. (1959): The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, **19**, 159-170.

[10] Downey, R.G. (1979): Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, **3**, 453-461.

[11] Espinosa, M.P and Gardeazabal, J. (2010): Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, **54**, 415-425.

[12] Espinosa, M.P. and Gardeazabal, J. (2013): Do students behave rationally in multiple choice tests? Evidence from a field experiment. *Journal of Economics and Management*, **9**, 107-135.

[13] Frary, R.B. (1989): Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, **2**, 79-96.

[14] García-Pérez, M.A. and Frary, R.B. (1991): Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology*, **44**, 45-73.

[15] Gifi, A. (1990): *Nonlinear Multivariate Analysis*. Chichester: Wiley.

[16] Greenacre, M. (2007): *Correspondence Analysis in Practice* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

[17] Guttman, L. (1941): The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed): *The Prediction of Personal Adjustment*, 321-345. New York: Social Science Research Council.

[18] Hendrickson, G.F. (1975): The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, **8**, 291-296.

[19] Lesage, E., Valcke, M., and Sabbe, E. (2013): Scoring methods for multiple choice assessment in higher education: Is it still a matter of number right

scoring or negative marking? *Studies in Educational Evaluation*, **39**, 188–193.

[20] Lord, F.M. (1958): Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, **23**, 291-296.

[21] Lord, F.M. (1975): Formula scoring and number-right scoring. *Journal of Educational Measurement*, **12**, 7-11.

[22] MATLAB R2012B [Computer software]. Natick, MA: The MathWorks.

[23] Nunnally, J. and Bernstein, I. (1994): *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

[24] Raffeld, P. (1975): The effects of Guttman weights on the reliability and predictive validity of objective Tests when omissions are not differentially weighted. *Journal of Educational Measurement*, **12**, 179-185.

[25] Reilly, R.R. and Jackson, R. (1973): Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement*, **10**, 185-194.

[26] Revuelta, J. (2005): An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika*, **70**, 305-324.

[27] Rodriguez, M.C. (2005): Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, **24**, 3–13.

[28] Sabers, D.L and Gordon, W. (1969): The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, **6**, 93-96.

[29] Slakter, M.J. (1967): Risk taking on objective examinations. *American Educational Research Journal*, **4**, 31-43.

[30] Sočan, G. (2009): Scoring of multiple choice items by means of internal linear weighting. *Review of Psychology*, **16**, 77-85.

[31] Ten Berge, J.M.F. (1993): *Least Squares Optimization in Multivariate Analysis.* Leiden: DSWO Press. Retrieved from http://www.ppsw.rug.nl/~kiers/leastsquaresbook.pdf

[32] Thissen, D. and Steinberg, L. (1984): A response model for multiple choice items. *Psychometrika*, **49**, 501-519.

# Appendix

**Table A:** Percentages of choices of various alternatives and metric properties of the number-correct score

| | | | Response | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Instruct. | k | Distractors | O | C | W1 | W2 | | $r_{\theta.NC}$ | $\alpha_{NC}$ |
| Forced choice | 15 | one positive | n/a | 33.4 | 34.2 | 32.5 | | .73 | .53 |
| | | both negative | n/a | 34.1 | 33.2 | 32.7 | | .79 | .63 |
| | 50 | one positive | n/a | 33.4 | 34.1 | 32.5 | | .89 | .79 |
| | | both negative | n/a | 34.0 | 33.2 | 32.7 | | .92 | .85 |
| | | Average | n/a | 33.7 | 33.7 | 32.6 | | .83 | .70 |
| Omissions allowed | 15 | one positive | 16.9 | 27.7 | 28.3 | 27.1 | | .70 | .59 |
| | | both negative | 16.7 | 28.3 | 27.7 | 27.3 | | .74 | .66 |
| | 50 | one positive | 16.9 | 27.7 | 28.3 | 27.1 | | .83 | .83 |
| | | both negative | 16.7 | 28.3 | 27.7 | 27.3 | | .85 | .86 |
| | | Average | 16.8 | 28.0 | 28.0 | 27.2 | | .78 | .74 |

*k* = number of items, O = omit, C = correct, W = wrong, $r_{\theta.NC}$ = validity of the number-correct score in the population, $\alpha_{NC}$ = coefficient alpha for the number-correct score in the population.

Metodološki zvezki, Vol. 12, No. 2, 2015, 45-68

# Web Mode as Part of Mixed-Mode Surveys of the General Population: An Approach to the Evaluation of Costs and Errors

Nejc Berzelak[1], Vasja Vehovar[2] and Katja Lozar Manfreda[3]

**Abstract**

Lower data collection costs make web surveys a promising alternative to conventional face-to-face and telephone surveys. A transition to the new mode has already been widely initiated in commercial research, but web surveys remains limited in academic and official research projects that typically require probability samples and high response rates. Various design approaches for coping with the problems of sampling frames, incomplete Internet use, and nonresponse in web surveys have been proposed. Mixed-mode designs and incentives are two common strategies to reach Internet non-users and increase the response rates in web surveys. However, such survey designs can substantially increase the costs, the complexity of administration and the possibility of uncontrolled measurement effects. This paper presents and demonstrates an approach to the evaluation of various survey designs with simultaneous consideration of the errors and costs. It focuses on the designs involving the web mode and discusses their potential to replace traditional modes for probability surveys of the general population. The main idea of this approach is that part of the cost savings enabled by the web mode can be allocated to incentives and complementary survey modes to compensate for the Internet non-coverage and the higher nonresponse. The described approach is demonstrated in an experimental case study that compares the performance of mixed-mode designs with the web mode and prepaid cash incentive with that of an official survey conducted using the face-to-face and telephone modes. The results show that the mixed-mode designs with the web mode and incentives can greatly increase the response

[1] Nejc Berzelak, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, 1000 Ljubljana; nejc.berzelak@fdv.uni-lj.si

[2] Vasja Vehovar, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, 1000 Ljubljana; vasja.vehovar@fdv.uni-lj.si

[3] Katja Lozar Manfreda, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, 1000 Ljubljana; katja.lozar@fdv.uni-lj.si

rate, which even surpasses that of the conventional survey modes, but still offer substantial cost savings. However, the results also show that higher response rate does not necessary translate to higher data quality, especially when the main aim is to obtain estimates that are highly comparable with those of the reference survey.

# 1 Introduction

Declining survey response rates and increasing survey costs force researchers to use new modes and procedures for survey data collection. Web surveys are one of the most promising approaches, especially in terms of cost savings and increased measurement quality (e.g. reduced social desirability bias due to the absence of interviewers, less mistakes in navigation or smaller item nonresponse due to the computerisation of the questionnaire, more effort from respondents who can choose to answer the questionnaire at the time and pace of their convenience, larger respondents' motivation due to the possibilities of including multimedia, etc.; Callegaro et al. 2015). Therefore, ESOMAR's reports on their prevalence in commercial research (ESOMAR, 2014) are not surprising. However, the breakthrough of web surveys in academic and official fields remains limited. This situation is mainly due to the lack of adequate sampling frames, incomplete Internet access and use in the general population and even lower response rates compared with the traditional modes (Lozar Manfreda et al., 2008; Shih and Fan, 2008). These drawbacks are particularly critical in probability samples, which are commonly required for academic and official data collection.

The potentials of web surveys to reduce research costs and increase measurement quality strongly encourage researchers to find the solutions for their application to the general population and to any other study with high data quality requirements, such as official statistic surveys. In this study, we examine a combination of two such solutions: incentives and mixed-mode survey designs. Incentives have been shown to be an effective means for increasing the overall response rates in traditional survey modes (Singer and Ye, 2013) and in web survey (Göritz, 2006). Mixed-mode designs aim at compensating for the weaknesses of each individual mode by concurrently or sequentially combining different modes within a single survey project (de Leeuw, 2005). In the case of web surveys, they can be used to reach Internet non-users and may also stimulate the response of Internet users who do not want to participate online.

This study mainly aims to present and demonstrate the methodology for benchmarking and evaluating the performance of a web survey included in a mixed-mode design with incentives in terms of 1) obtaining the results comparable to those from traditional survey modes and 2) reducing the overall survey costs in comparison with traditional modes. We put a special emphasis on the issues of Internet non-coverage and nonresponse, which are the most typical and specific problems of web surveys.

We begin with an introductory overview of the use of mixed-mode approaches and incentives to deal with the problems of incomplete Internet use and

nonresponse in web surveys. In the second part, we explain a methodological approach for the evaluation of different survey designs in terms of errors and costs. We establish the criteria for the identification of the optimal design and apply them to the empirical data from a case study. Finally, we present and discuss the results of the case study by observing the differences in sample composition, substantial data and costs.

# 2 Background

The problem of incomplete Internet access and use is one of the greatest threats to inference from web surveys of the general population. The differences in Internet use among countries are profound. In the European Union (EU), the proportion of Internet users ranges from 54% in Romania to 96% in Denmark (Eurostat, 2015). At the world level, according to Internet World Stats (2014), the proportion of Internet users in 2013 and 2014 was below 30% in Africa and almost 88% in North America. Only a few countries have Internet coverage above 90%, and in Europe these countries are Switzerland, the United Kingdom, Sweden, Finland, the Netherlands, Luxembourg, Norway, Denmark and Iceland (Eurostat, 2015).

The bias due to Internet non-coverage occurs if Internet non-users in the target population differ from Internet users in the characteristics measured by survey questions. Internet non-users are typically older, less educated, live in lower-income households, work in manual labour or are unemployed (Eurostat, 2015). The differences between users and non-users can contribute to the Internet non-coverage bias when a web survey is applied to the general population (Couper et al., 2007; Dever et al., 2008; Rookey et al., 2008). Although attempts to infer to the general population using post-survey adjustments have been made, their performance is often questionable and inconsistent across different variables and surveys (Lee and Valliant, 2009; Loosveldt and Sonck, 2008; Schonlau et al., 2009).

Low response rates are another prominent problem of web surveys. Whereas a persistent trend of declining response rates is observed in all survey modes (de Leeuw and de Heer, 2002), the situation is even worse in web surveys, which have been found to commonly produce lower response rates than the other modes (Lozar Manfreda et al., 2008; Shih and Fan, 2008).

Both Internet non-coverage and nonresponse problems can critically compromise data quality in web surveys by failing to reach specific parts of the target population. In consideration of the advantages of online data collection, particularly reduced costs and speed, significant efforts are unsurprisingly being devoted to the development of designs for a more efficient implementation of web surveys on the general population.

The two most plausible solutions for these problems are incentives to stimulate participation and mixed-mode survey designs that use an alternative mode to survey sampled individuals not reached by the web mode. An alternative approach

of online probability-based panels, which can also provide Internet access to households without the Internet, is clearly expensive. Consequently, only a few such panels currently exist, including LISS panel in the Netherlands, GESIS and GIP panels in Germany, ELIPSS in France, Social Science Research Institute panel in Iceland, and GfK-Knowledge Network panel in the United States.

Therefore, we limit our discussion to the more generally feasible mixed-mode survey designs with incentives in this study. Although past studies have already devoted attention to these topics also in the context of web surveys, the elaboration of cost-related factors is severely lacking. Thus, we first present the main general points related to mixed-mode designs involving web surveys and incentives, and then elaborate and demonstrate an approach for the evaluation of costs and errors in such designs.

## 2.1 Mixed-mode survey designs

Although mixed-mode survey designs existed well before the emergence of web surveys, the specifics of the new mode strongly facilitated their development and use. The main rationale for using mixed-mode designs involving the web mode is to exploit the cost advantages of web surveys and at the same time overcome their main problems, particularly non-coverage and nonresponse.

Modes can be combined in various ways at all stages of the survey project: at the initial contact stage, during data collection (response) stage or during the follow-up of non-respondents (de Leeuw, 2005). Many mode combinations are possible because the contact and data collection stages of web surveys are explicitly separated (Vehovar and Lozar Manfreda, 2008). The three most common combinations are as follows:

1. Telephone, mail or even SMS **invitation to a web survey** to overcome the problem of lacking contact information in the sampling frames and to stimulate the response rates by increasing the legitimacy of the request (e.g. Bosnjak et al., 2008; Messer and Dillman, 2011; Porter and Whitcomb, 2007).
2. **Concurrent use of web and other data collection modes** for different sample members at the same stage of the data collection to overcome the Internet non-coverage problem and to increase response. The most appropriate mode for each respondent can be selected in advance by the researcher (Blyth 2008; Rookey et al., 2008) or offered as a choice to the respondent (e.g. Medway and Fulton 2012; Parackal, 2003; Vehovar et al., 2001). However, although the latter option may seem respondent-friendly and advantageous, there is a strong evidence that offering a choice often decreases rather than increases the response rate (Dillman et al., 2014; Medway and Fulton 2012). This approach is therefore generally not recommended.

3. **Sequential use of different data collection modes** to increase the response rates and to reach Internet non-users. The most common approach is to begin with the web mode and then follow up the web non-respondents using one of the traditional survey modes (e.g. Dillman et al., 2014; Greene et al., 2008; McMorris et al., 2009). Alternatively, the web can be used to follow up non-respondents that used the other modes (e.g. Greene et al., 2008; McMorris et al., 2009).

Mixed-mode designs have specific problems. First, their administration is usually significantly more complex than that of single-mode surveys. A well-established sample monitoring system is essential to assure a proper assignment and transition of sampled individuals to different modes. The additional workload and required resources also have a direct effect on the costs of a survey project.

Second, the different modes may affect the comparability of the data obtained. Each mode has specific characteristics that may influence the answers provided by respondents. For example, compared with telephone and face-to-face surveys, the web mode is self-administered rather than interviewer-administered, the questions are presented visually rather than aurally, and the responses are provided by electronic input rather than orally. To what degree these factors cause mode effects and the related between-mode differences remains to be thoroughly investigated, as the findings of methodological studies are currently largely inconsistent. Some of these studies have found no differences among the modes (Knapp and Kirk 2003; Revilla and Saris 2010), whereas others have reported mode effects in different data quality indicators. Examples of mode effects include between-mode differences in the response order effects (e.g. Galesic et al., 2008; Malhotra 2008), response length to open-ended questions (Kwak & Radler, 2002), non-substantive answers (Bech and Kristensen, 2009), item nonresponse (Heerwegh and Loosveldt, 2008; Smyth et al., 2008) and non-differentiation (Tourangeau, 2004; Fricker et al., 2005). The most consistently observed effects are related to social desirability, in which the web mode almost universally produces less socially desirable answers than the interviewer-administered modes (e.g. Lozar Manfreda and Vehovar, 2002; Kreuter et al., 2008; Tourangeau et al., 2013).

Empirically investigating the mode effects is a demanding task that requires a researcher to disassemble error sources into individual components (sampling, non-coverage, nonresponse and measurement errors) and to identify which of these components are affected by the mode. This process is usually only possible with complex and well-controlled experimental designs. However, when the main objective is to compare the overall performance of different survey designs, all error sources of each individual mode can be taken into account without the separate identification of each component. We use the latter approach in this paper, as our goal is not to analyse different error sources but to observe the overall comparability of the obtained estimates among survey designs.

## 2.2 Incentives

Based on the evidence from research on traditional survey modes (Church, 1993; Singer and Ye, 2013), several studies have expected and confirmed that incentives can also improve the response rates in web surveys (Göritz, 2006; Parsons and Manierre, 2014). However, the efficiency of the incentives strongly depends on their type (monetary or non-monetary, pre-paid, post-paid, or lottery), value, survey sponsor (commercial or non-commercial), sample type (panel or non-panel studies) and several other factors (see e.g. Bosnjak and Tuten, 2003; Downes-Le Guin et al., 2002; Göritz, 2006, 2008; Göritz et al., 2008).

Pre-paid incentives have proved to be more effective than post-paid (promised) incentives for different survey modes (e.g. Church, 1993; James and Bolstein, 1992). Their benefits are even higher in web surveys because they increase the legitimacy of the survey request that is often limited online. Furthermore, immediate delivery dispels respondents' doubts about whether the incentive will actually be delivered or not. A small-value pre-paid incentive can already significantly improve the response rate, with higher values bringing relatively small additional benefits (e.g. Downes-Le Guin et al., 2002; Villar, 2013). However, the use of pre-paid incentives in web surveys is mostly limited to list-based surveys, in which people are personally addressable prior to their participation in the survey.

Although post-paid incentives are usually less effective, Göritz (2006) suggested that they seem to work better in online rather than in offline studies. A possible reason is that such incentives are common online and that Internet users may have gotten used to them (Bosnjak and Tuten, 2003).

Despite the potential benefits to the response rate, incentives may increase the nonresponse bias by attracting specific sample members (e.g. people with lower income, with more free time, who are younger, who are more computer literate, etc.). Göritz (2008) showed that incentives had a greater effect on low-income members of an online panel. Altered sample compositions have also been reported also by some other authors (e.g. Parsons and Manierre, 2014; Sajti and Enander, 1999). Furthermore, incentives may stimulate some respondents to complete the questionnaire more than once. This scenario is most likely to happen if a web survey is not based on a list of sampled persons where the possibilities for effective access control are limited (Comley, 2000).

Incentives are usually expected to increase the commitment of respondents to the task of answering questions because they feel compensated for their effort. However, with post-paid incentives, some people may use non-optimal response strategies just to reach the end of the questionnaire and become eligible for the incentive. Little empirical evidence about the effects of incentives on data quality exists. Göritz et al. (2008) found generally high data quality without systematic differences in the item nonresponse, length of answers to open-ended questions, discrepancy of answers, completion time and response styles between respondents receiving and those not receiving the incentive.

Finally, the effect of incentives may depend on the whole context of a specific web survey. A meta-content analysis of a larger number of web surveys (Lozar Manfreda and Vehovar, 2002) showed that incentives decreased the drop-out rates but did not increase the proportion of respondents who began the survey participation. The latter was more influenced by other design features, such as pre-notification and content of the questions. The meta-analysis by Cook et al. (2000) found even lower response rates with incentives, although the authors attributed this outcome to the cofounding between survey length and use of incentives, where disproportionately long or tedious surveys more often offered incentives.

### 2.3 Consideration of costs and errors

Most of the existing studies that compare web surveys with other survey modes have focused on the response rates (Dolničar et al., 2009; Lozar Manfreda et al., 2008; Shih and Fan, 2008). These comparisons are insufficient, as the response rate is the only one indicator of data quality used. Existing comparisons also rarely consider the costs of data collection, which can be crucial for an overall assessment of web survey performance. That is, the cost savings earned by using a web survey instead of a more expensive survey mode (e.g. face-to-face or telephone) can be invested into additional recruitment measures (e.g. incentives) to increase the response rates. Such measures can significantly improve the performance of the web mode.

The costs–error evaluation of different alternative survey designs requires the separate consideration of costs at all contact and data collection stages for each mode. Several means can be used to identify the optimum approach among competing survey designs by taking into account the errors and costs. Some strategies include finding the design with the lowest costs at the fixed error level, the design with the smallest error at fixed costs (available budget), or the design with the smallest product of costs and a selected error measure (Vehovar et al., 2010).

These calculations rely on certain assumptions regarding the errors that are expected to occur in the actual survey implementation. Informed decision making about these assumptions can be made on the basis of experience, similar surveys or a pilot study.

## 3 Case study comparing errors and costs: methodology

To explore the possibilities of surveying the general population using the web mode, we performed an experimental study as part of a survey on information communication technologies (ICT) in Slovenia in 2008. According to Eurostat, 77% of Slovene households had Internet access in 2014, although this percentage was around 60% at the time of the survey. This penetration rate is too low to make

a web survey feasible as a standalone mode for surveying the general population. To overcome the Internet non-coverage problem, we used a mixed-mode design. We also used incentives to address the problem of nonresponse, which was expected to be high.

## 3.1 Experimental design

The Survey on ICT Usage by Individuals and in Households is a Eurostat survey conducted in all EU member states on an annual basis. In Slovenia, this survey is conducted by the Statistical Office of the Republic of Slovenia (SORS). In April and May 2008, the SORS fielded the survey using the mixed-mode design with face-to-face and telephone modes, depending on the availability of telephone numbers (further referred to as the *SORS survey*). The sample of 2,504 Slovenian citizens, together with their home addresses, was obtained from the Central Register of Population (CRP). Parallel to the official SORS survey, we implemented several experimental mixed-mode designs based on the same questionnaire (Berzelak et al., 2008).

In this study, we focused on the experimental mixed-mode designs with a mail follow-up of web survey non-respondents and two incentive conditions. In June 2008, mail invitations to take part in the web survey were sent to 305 randomly sampled individuals from the CRP (further on *"web > mail" design*). The letters included questionnaire access instructions, a unique access code for each sampled person and the statement that non-respondents would receive a paper questionnaire in a follow-up letter. The follow-up was carried out 10 days later. Two weeks later, the second follow-up mail letter was sent to the remaining non-respondents. Access to the web questionnaire was available throughout the whole data collection period.

The sampled individuals were randomly assigned to two incentive treatments: 100 sampled members received a EUR 5 banknote as a pre-paid incentive in the first mail letter and the remaining 205 received no incentive.

The survey questionnaire consisted of 44 questions (approximately 125 items) covering the topics like household access to various ICTs, individual's frequency of computer and Internet use, devices used to access the Internet, frequency of using various online services, experiences with online shopping, and background information about the target person.

## 3.2 Research questions

On the basis of the experimental data, we explored the following research questions:

> **Q1. Response rate**: Can the "web > mail" mixed-mode design with pre-paid cash incentives produce a response rate comparable with that in the reference SORS survey?

**Q2. Sample composition**: How do the final sample compositions of the "web > mail" designs with and without incentives differ from the reference SORS data?

**Q3. Difference in substantive results**: What is the difference between the survey estimates of the "web > mail" designs with and without incentives compared to the reference SORS data?

**Q4.** Does **weighting** based on demographic variables eliminate any differences in estimates between the "web > mail" designs and the SORS survey?

**Q5. Costs**: What is the cost of the compared designs for equal target sample sizes?

**Q6. Cost and error comparison**: Which "web > mail" design (with or without incentives) performs better when considering the costs and errors simultaneously?

To investigate these research questions, we performed several comparisons between the experimental survey designs and the official SORS survey.

## 3.3 Estimation of the bias

Although the **response rates** are straightforward to calculate, they have been shown to be insufficient predictors of the nonresponse bias (Groves & Peytcheva, 2008). Thus, we also compared **sample composition** between the experimental designs and the SORS survey, considering the final sample composition and the sample composition gained prior to the mail follow-up, which includes only the Internet users.

To **compare substantive results**, we selected key survey variables associated with ICT use. We compared the unweighted and weighted estimates from the experimental designs with the weighted reference SORS survey. The weights for the experimental survey designs were calculated using the raking method by sex, age and educational structure of respondents, as it is usually also done by the SORS. As the differences in estimates can substantially depend on the content of the variables, we also calculated the average difference in substantive results across several variables.

To calculate the difference in estimates for each variable, we followed the definition of the bias as the difference between the expected value over all respondents and the true value of the estimated parameter (Groves et al., 2004):

$$Bias(\bar{y}) = E(\bar{y}) - \bar{Y} \tag{3.1}$$

The above definition of the bias considers all the error sources that may affect the differences between the compared modes, including Internet non-coverage, nonresponse and measurement errors. Although it does not enable the explicit separation of error sources, it suffices for the evaluation of the overall differences among the compared survey designs.

Clearly, the true value is rarely known in practice and is usually estimated using a reference ("gold standard") survey. In such cases, the above equation can be extended to account for the variance in the reference survey estimates (Biemer, 2010). However, because our main goal is to assess to what degree the experimental survey designs can provide results comparable with the official survey, we regard the estimates from the SORS survey as true values. Therefore, we refer to the difference between the experimental designs and the SORS survey as the bias, and it is calculated using Equation (3.1).

As we want to estimate the bias across several survey items and compare them among the experimental designs, we calculated the standardised bias for each item. The standardised bias is defined as the absolute bias divided by the standard deviation of the item in the reference SORS survey:

$$St.bias(\bar{y}) = \frac{E(\bar{y}) - \bar{Y}}{SD(\bar{Y})} \ . \tag{3.2}$$

For the items measured on the nominal scale, the statistic of interest is the proportion, and the standardised bias for each item is calculated as follows:

$$St.bias(p) = \frac{p - \pi}{\pi(1-\pi)} . \tag{3.3}$$

We calculated the average standardised bias across items in each experimental design for each data collection stage (before and after the mail follow-up) and for the unweighted and weighted data. This approach enables the evaluation of both the incentives and the mixed-mode follow-up. Note that we did not test the statistical significance of the observed differences (biases) among the modes, as the objective was not to make inference about the target population but to investigate the presence of the bias in the specific survey implementation compared with the reference survey.

## 3.4 Identification of the optimal design

The identification of the most optimal survey design requires a simultaneous evaluation of costs and errors. To estimate the overall costs of each design, we considered the fixed and variable costs at each contact and data collection stage. We then included them in the cost model outlined by Vehovar et al. (2010) to simulate the costs at various target sample sizes.

When looking for the most optimal survey design in terms of errors and costs, it is beneficial to consider both, the bias of the estimate and also the corresponding sampling variance $Var(\bar{y})$. Namely, the variance decreases with the sample size *n*. If the lower costs of a specific survey design enable the increase in the sample size, the sampling variance of an item of interest will be reduced. Following the *total survey error* principles, each item can be evaluated using the standard approach of the root mean squared error (RMSE) (Kish, 1965):

$$RMSE(\bar{y}) = \sqrt{Var(\bar{y}) + Bias(\bar{y})^2} \,. \tag{3.4}$$

The total survey error does not depend only on the difference between the true value and the expected value from the survey but also on the precision of the point estimates. Therefore, the potential higher bias of a cheaper survey mode may be outweighed by the lower sampling variance enabled by the larger sample size in some circumstances.

Similar to the bias, the RMSE can be aggregated across several survey items. The relative RMSE is obtained for each item by dividing the RMSE by the value obtained from the reference SORS survey, against which we compare the estimated parameters:

$$rRMSE(\bar{y}) = \frac{\sqrt{Var(\bar{y}) + Bias(\bar{y})^2}}{\bar{Y}} \,. \tag{3.5}$$

Thus, we were able to calculate the average rRMSE across the selected key survey items. To evaluate the costs–error performance of each design, we observed which of the two experimental designs would produce the smallest average rRMSE at the given fixed budget. This circumstance resembles the practical situation survey researchers often face when deciding on how to best spend the available budget to reach the highest possible data quality, which can be defined in terms of accuracy of estimates, comparability with another survey (as in our case) or many other quality-related criteria. To simulate the amount of the total error in the two "web > mail" designs, we assumed that the difference between the designs and the official SORS survey would remain unchanged, but the sampling variance would change because of the different sample sizes reachable by the specified budget.

# 4 Results

## 4.1 Response rates

Comparing the response rates in the experimental mixed-mode designs ("web > mail" designs) with those in the official SORS survey enables us to explore whether such designs can produce response rates comparable with those in the traditional modes. We expect follow-up stages in the mixed-mode design to increase the response rates by reaching the sampled Internet non-users and converting those who do not want to participate online for some reason. We expect additional increases in the response rates in the experimental group with the pre-paid incentive.

The unit response rates for the compared survey designs by individual stages of the survey project are presented in Figure 1. The response rates are calculated as the percentage of the survey respondents out of all the eligible sampled individuals.

The **mixed-mode design without incentives** produced a final response rate of 33%. At the first stage, when the respondents were only able to complete the online questionnaire, only 11% of the sampled individuals participated in the survey. The follow-up reminder with paper (mail) questionnaire increased the response rate by 13 percentage points, and the final reminder increased it by another 9 percentage points. It is possible that some of the respondents would have completed the web questionnaire later even without follow-up. However, this proportion is unlikely to be high, considering that only 21% of those who participated after the second and the third reminders opted for the web mode. These findings confirm the result of previous research that offering a follow-up in a different mode can help increase the response rates.

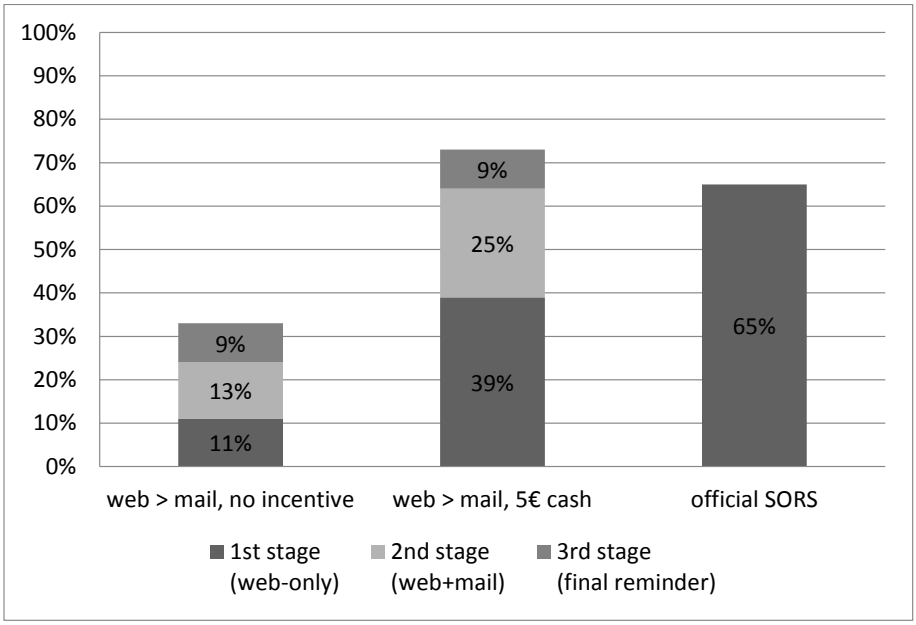



**Figure 1:** Unit response rates produced by the evaluated survey designs.

The observed **effect of incentives** on the response rates in the mixed-mode survey design is profound. Figure 1 shows that the EUR 5 pre-paid cash incentives already increased the response rate from 11% to 39% at the first stage, in which the respondents were able to complete the online questionnaire only. Although the effect somewhat diminished at the follow-up stages, it still produced the final response rate of 73%, which surpassed the response rate achieved by the official SORS survey. This finding confirms the large positive effect of pre-paid cash incentives on increasing the willingness of participation in both web and mail modes.

Considering only response rates, the results indicate that the combination of web and mail modes with appropriate incentives may be used to replace the traditional combination of face-to-face and telephone modes to survey the general population. However, as noted earlier, the response rate should not be regarded as

the only indicator of data quality because it is not necessarily a good predictor of the nonresponse bias.

## 4.2 Composition of the samples

We explored the biases in the sample composition of the experimental "web > mail" designs compared with the official SORS survey. The weighted official data were used as reference, as weighting can be expected to bring the compared socio-demographic estimates very close to the population characteristics.

Note that various error sources, including sampling frame non-coverage, Internet non-coverage, nonresponse and measurement errors, can contribute to a biased sample composition. As all evaluated survey designs use the same highly accurate sampling frame of the CRP, the errors due to sampling frame non-coverage are likely to be low. The same is true for the between-mode differences in measurement errors because the analysed questions are simple, factual, and not explicitly sensitive. Finally, the problem of Internet non-coverage is avoided in the "web > mail" designs by enabling Internet non-users to complete the paper questionnaire in the follow-up stages. Therefore, we expect that any bias in the final sample composition of the "web > mail" design compared with that in the official SORS survey is mostly due to the differences in the nonresponse error. However, the biases in the first (web-only) stage of the experimental designs are likely to be due to both nonresponse and Internet non-coverage.

The reduction of Internet non-coverage problem can be observed by comparing the estimates across the stages of the "web > mail" designs. At the first stage (web-only) of the design without incentives, the percentage of daily Internet users was 91%. After the mail option was offered (2$^{nd}$ stage) and final reminders sent (3$^{rd}$ stage), this proportion dropped to the final 62%. Although the proportion of Internet users did not reach the official estimate of 56% in the SORS survey, it substantially reduced the differences compared to the first (web-only) stage. While the web mode at the first stage of the design with incentives was more successful in reaching less frequent Internet users compared to the design without incentives (82% reported using the Internet almost every day), the decrease after the mail follow-up was much less pronounced and produced the final estimate of 73% of daily Internet users

Consistent with previous research, Table 1 shows that the **sample composition produced at the first stage** of the "web > mail" design without incentives is substantially affected by the typical characteristics of Internet users. Compared to the official survey, the sample over-represents men, younger and those with higher income. However, the education of respondents is similar to the official data. With the incentives, the respondents of the first stage of the "web > mail" design are similar to those of the SORS survey in terms of gender and age, but the bias for education is larger. This finding suggests that incentives especially attract female, older and higher-educated Internet users.

The **final sample composition** after the mail follow-ups in the "web > mail" design without incentives shows the increased proportion of women, older respondents and respondents with lower income. All of these characteristics are similar to those of the official SORS survey, although the average age remains substantially lower. This finding indicates that more respondents with characteristics consistent with Internet non-users were reached by the follow-ups, thus resulting in decreased non-coverage error. However, the combination of mail follow-up and web survey with pre-paid incentives produced mixed results. The mail follow-up made the sample more similar to that of the SORS sample in terms of age, income and education, but it increased the bias in the gender structure.

**Table 1:** Composition of samples produced by the evaluated survey designs.

| | Web > mail (unweighted) | | | | Official SORS |
|---|---|---|---|---|---|
| | No incentive | | €5 cash | | |
| | Stage 1: Web-only (n=22) | Final: Web+mail (n=67) | Stage 1: Web-only (n=39) | Final: Web+mail (n=73) | (n=1052) |
| % of men | 64% | 52% | 49% | 45% | 52% |
| Mean age | 24.7 | 29.0 | 27.9 | 28.3 | 32.5 |
| Mean household size | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 |
| Median household month. income (EUR) | 2,000 | 1,500 | 2,000 | 1,700 | 1,500 |
| % with more than secondary education | 20% | 21% | 28% | 24% | 19% |

The findings suggest that although the web survey with pre-paid cash incentives combined with the mail questionnaire significantly increased the response rates, it did not produce the sample composition similar to that in the SORS survey. The biases were especially large for gender, age and education. Interestingly, despite the profoundly lower response rates, the "web > mail" design without incentives produced the basic characteristics of respondents more similar to the official data.

## 4.3 Biases in substantive items

The socio-demographic composition of a sample can be largely corrected by weighting if the relevant information is available from reliable official sources, as in our case. To gain further insights into the performance of the compared experimental designs, we analysed the biases in substantive variables related to the survey topic using unweighted and weighted data. Again, bias is defined as the difference in estimates between the experimental designs and the official data, as

the comparability of estimates is considered to be the most important data quality criterion in this case.

In the unweighted data, the differences between the "web > mail" designs and the SORS data are higher in the "web > mail" design with incentives than in that without incentives for six out of the seven selected key survey variables (Table 2). The opposite situation is observed only for the question about online shopping. In both experimental designs, the biases are most profound in the estimates related to computer and Internet use. The situation does not improve much after weighting by sex, age and education. Although the differences between the "web > mail" designs and the official survey are reduced for most variables, the improvement is relatively limited. For most variables the biases remain higher in the design with incentives.

**Table 2:** Biases as the differences in estimates between the experimental designs and the official data for selected variables before and after weighting.

| | Web > mail (final) | | | | Official SORS |
|---|---|---|---|---|---|
| | No incentive | | €5 cash | | |
| | Unweighted | Weighted | Unweighted | Weighted | (n=1052) |
| Mean number of ICT devices in household[a]) | 4.13 (-0.08) | 4.15 (-0.06) | 4.47 (+0.26) | 4.50 (+0.29) | 4.21 |
| % living in household with Internet access | 86% (+4 pp) | 87% (+5 pp) | 93% (+11 pp) | 91% (+9 pp) | 82% |
| % living in household with broadband Internet access | 70% (-1 pp) | 70% (-1 pp) | 73% (+2 pp) | 69% (-2 pp) | 71% |
| % using computer every day or almost every day | 76% (+15 pp) | 72% (+11 pp) | 83% (+22 pp) | 80% (+19 pp) | 61% |
| % using internet every day or almost every day | 62% (+6 pp) | 56% (+0 pp) | 76% (+20 pp) | 73% (+17 pp) | 56% |
| % using mobile telephone | 100% (+3 pp) | 100% (+3 pp) | 100% (+3 pp) | 100% (+3 pp) | 97% |
| % shopping over the Internet in the last 3 months | 28% (+13 pp) | 26% (+11 pp) | 24% (+9 pp) | 22% (+7 pp) | 15% |

[a] The possession of the following devices was counted: TV, landline telephone, mobile telephone, desktop computer, laptop computer, personal digital assistant and gaming console.

Next, we calculated the average standardised bias in the estimates of the proportions in the "web > mail" designs against the official SORS survey for all 117 categorical variables in the questionnaire (Table 3). We did not include the comparison of mean estimates because only six items were measured at the scale level and even those were related more to background information than substantive

topics. The included variables thus take into account all key aspects of ICT use covered by the questionnaire (frequency of computer and Internet use, devices, and online services).

The unweighted data after the first web-only stage show that the average difference between the "web > mail" designs and the official survey is lower in the design with incentives (1.36) than in the design without incentives (1.51). However, the opposite is observed after the mail follow-up stages. Although offering the mail option decreased the average difference in both "web > mail" designs, the improvement was higher in the design without incentives (0.95 vs. 1.04). The weights further slightly decreased the average differences, but the design with incentives again performed worse. Note that the weights were only calculated for the final data. This is primarily due to small sample sizes after the first stage, which limited the possibilities to apply weighting procedures comparable to those used by the statistical office.

**Table 3:** Average standardised biases in the estimates of proportions as the differences between the experimental designs and the official data.

| | Web > mail | | | |
|---|---|---|---|---|
| | No incentive | | €5 cash | |
| | Stage 1: Web-only (n=22) | Final: Web+mail (n=67) | Stage 1: Web-only (n=39) | Final: Web+mail (n=73) |
| Average standardised difference for proportions (unweighted data) | 1.51 | 0.95 | 1.36 | 1.04 |
| Average standardised difference for proportions (weighted data) | n/a | 0.87 | n/a | 0.99 |

We can conclude that monetary pre-paid incentives can significantly increase participation in a mixed-mode design involving the web mode. The response rates in our case even surpassed those using the traditional telephone and face-to-face combination without incentives. However, the incentives strongly influenced the self-selection of specific respondents and increased the bias, calculated as the difference in estimates against the official data. This finding suggests that the mail follow-up is more effective in reducing the difference between the experimental survey designs and the official survey than incentives.

## 4.4 Evaluation of costs and errors

As we stressed above, the consideration of research costs is crucial to understand the performance of a specific survey design. This is especially true for web surveys, in which cost savings may enable larger sample sizes and implementation of additional measures to increase the data quality.

To illustrate the cost advantages and to perform further comparisons, we calculated the costs of obtaining a hypothetical sample size of 1,000 respondents for each evaluated survey design, taking into account the response rates obtained in the presented case study (Figure 1). Table 4 shows that the two "web > mail" designs are substantially cheaper than the combined face-to-face and telephone survey. This is mainly due to the elimination of the interviewer-related costs. Comparing both "web > mail" designs, the costs are clearly substantially higher in the design with the pre-paid incentives.

**Table 4:** Costs of data collection for a simulated final sample size of n=1.000 in the compared survey designs.

| | Web > mail | | | | | | Official SORS |
|---|---|---|---|---|---|---|---|
| | No incentive | | | €5 cash | | | |
| | Stage 1 | Stages 1 & 2 | Stages 1, 2 & 3 | Stage 1 | Stage 1 & 2 | Stages 1, 2 & 3 | |
| Cost for the final sample size n = 1000 | €1,835 | €5,418 | €6,857 | €8,206 | €9,919 | €10,374 | €25,885 |

To demonstrate the approach for identifying the more optimal design, we searched for the design with the lowest average relative RMSE for a given fixed budget. Note that the RMSE calculation does not apply to the official SORS survey, used as reference with which the experimental survey designs are compared. As stated above, our study explores the most optimal alternative design to provide the results as comparable to the existing SORS survey as possible, but with potentially lower costs. The key data quality indicator is in this case comparability rather than accuracy. Correspondingly, the estimates from the SORS survey are considered to be "true values" against which the alternative designs are compared, making the RMSE calculation for the SORS survey inapplicable.

Table 5 presents the sample sizes that can be reached at a fixed budget of EUR 10,000 for the three compared survey designs. Assuming that the response rates are the same as in our empirical case, larger initial and final sample sizes can be achieved by the "web > mail" survey designs than by the face-to-face/telephone survey.

As noted earlier, the simulation of the total error in the two "web > mail" designs at the given fixed budget of EUR 10,000 assumes that the biases remain equal as in the empirical case, but considers the changes in the sampling variance because of the different sample sizes. Table 5 shows that the average relative RMSE across the seven key variables at a fixed budget is lower for the "web > mail" design without incentives than that with incentives. In consideration of our previous findings, this result is not surprising as the design with incentives produced greater differences with respect to the reference survey (Tables 1–3) and is substantially more expensive to implement (Table 4).

**Table 5:** Comparison of the mean RMSE for the selected seven key variables at a fixed budget of EUR 10,000.

| Survey design | Initial sample size | Final response rate | Final sample size | Mean rRMSE | rRMSE × costs |
|---|---|---|---|---|---|
| Web > mail, no incentive | 4,610 | 33% | 1507 | 0.8272 | 8272.74 |
| Web > mail, €5 cash | 1,318 | 74% | 962 | 0.9077 | 9078.47 |
| Official SORS | 580 | 65 % | 372 | n/a | n/a |

This demonstration adds another perspective to the above findings that a higher response rate influenced by incentives does not provide more comparable data to the official face-to-face/telephone survey. Again, this finding emphasises the inconsistent relation between response rates and the nonresponse bias. That is, higher response rates do not necessarily translate to lower nonresponse bias.

# 5 Summary and conclusions

The answer to the question of whether web surveys can be a viable alternative to traditional modes for surveys of the general population is not simple and straightforward. It requires a consideration of several issues and a careful planning of measures, including the evaluation of survey design elements to deal with Internet non-coverage and nonresponse problems. Furthermore, selecting and applying appropriate criteria is necessary to compare the costs and errors among different survey designs to find the optimal solution for specific research needs. This is particularly relevant to web surveys because of their substantial cost-saving potential. Of course, the magnitude of errors, costs of survey design implementation, and definition of the optimal ratio between costs and errors strongly depend on a specific survey project. This study mainly aimed to discuss an approach for the simultaneous evaluation of costs and errors to assess various survey design possibilities.

We discussed mixed-mode designs involving the web mode and incentives as two common measures to treat the Internet non-coverage and nonresponse problems of web surveys. In the survey practice, a web survey is usually considered as a complimentary or a replacement mode for an existing face-to-face or telephone survey. Comparability of data between the two modes and sufficient cost savings to justify the change in survey design are usually the most important evaluation criteria in such situations. As demonstrated, the simultaneous observation of response rates, biases (in our case defined in terms of the difference against the official survey), RMSE across several key variables and costs can produce a much higher informative value about the design's performance than response rates as the sole indicator of the data quality.

In our case study, the provision of EUR 5 cash incentives had a tremendous effect on the response rate in the design with a mail follow-up to the web survey. It increased the survey participation by 40 percentage points compared to the design without incentives. The response rate even surpassed that of the official survey that combines the face-to-face and telephone modes.

However, we identified several differences in the substantive results between the mixed-mode design with the web mode and the official face-to-face/telephone survey. Although weighting generally decreased the differences between the modes, they remained substantial in some variables. Furthermore, despite the substantially higher response rate, the design with incentives led to higher differences in estimates. This finding again confirmed that response rates are not necessarily good predictors of the nonresponse bias (Groves & Peytcheva, 2008).

We also demonstrated that the web mode could substantially decrease survey costs unlike a combined face-to-face and telephone survey. This finding remains true even when the web mode is included as part of the mixed-mode design rather than as a standalone mode and when pre-paid incentives are provided to all sampled persons. However, the reduced sampling variance, which resulted from the increased sample size obtained due to lower costs, was not sufficient to substantially reduce the total survey error, which was defined in terms of the comparability of estimates with those of the official survey. We showed that the relative bias was greater in the case of incentives, which strongly influenced the self-selection of specific respondents. Thus, the total survey error even increased despite the higher response rate. Incentives should therefore not be regarded as a universal tool for data quality improvements. It is important to critically evaluate their appropriateness in the context of a specific survey project and assess their performance as part of the survey pretesting.

In what follows, we present the limitations of the presented case study and the application of the proposed methodology for the evaluation of costs and errors. The sample sizes in the two experimental mixed-mode designs were small, potentially leading to unstable estimates. Furthermore, our calculations of the biases assumed that the official survey provided the true values of the estimated parameters. This issue is not necessarily a limitation by itself, as the comparability between modes is sometimes even a more important aspect of data quality than accuracy. This condition is especially true when the aim is to prevent breaks in time series, as is the case with many longitudinal surveys. Nevertheless, determining which design provides the estimates closer to the true population value would be certainly beneficial. Moreover, the evaluated designs presented only a small subset of the possible mixed-mode implementations. Alternative designs, such as simultaneous mode options or other mode combinations, could result in different relations between errors and costs.

It is also important to take into account that the study covers only one and in this context relatively specific survey topic, i.e. ICT usage. The survey topic is directly correlated with the Internet use as the key prerequisite to participate in a web survey, which was offered in the first stage of the investigated survey designs.

Furthermore, the survey questionnaire contained predominantly factual rather than attitudinal questions. Repeating the study with different survey topics would be informative as other topics my result in different biases caused by self-selection and mode effects.

Despite these limitations, the discussed methodology offers survey researchers a valuable tool for making informed decisions on the feasibility of various survey design possibilities. The proposed approach can provide important insights when a switch to alternative data collection modes is considered for an existing survey. The presented case study also confirms the efficiency of mixed-mode designs and incentives to improve the response rates in web surveys, but at the same time it cautions against the over-reliance on response rates in data quality assessment.

# References

[1] Bech, M. and Kristensen, M. B. (2009): Differential response rates in postal and web-based surveys among older respondents. *Survey Research methods*, **3**, 1-6.

[2] Berzelak, N., Vehovar, V., and Lozar Manfreda, K. (2008): *Nonresponse in web surveys within the context of survey errors and survey costs.* Paper presented at the 2nd MESS Workshop, Zeist, The Netherlands.

[3] Biemer, P. P. (2010): Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, **74**, 817-848.

[4] Blyth, B. (2008): Mixed-mode: The only "fitness" regime? *International Journal of Market Research*, **50**, 241-266.

[5] Bosnjak, M. and Tuten, T. L. (2003): Prepaid and promised incentives in web surveys: An experiment. *Social Science Computer Review*, **21**, 208-217.

[6] Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., and Kaczmirek, L. (2008): Prenotification in web-based access panel surveys: The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review*, **26**, 213-223.

[7] Callegaro, M., Lozar Manfreda, K., and Vehovar, V. (2015): *Web survey methodology*. Los Angeles [etc.]: Sage.

[8] Church, A. H. (1993): Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, **57**, 62-79.

[9] Comley, P. (2000): *Pop-up surveys. What works, what doesn't work and what will work in the future.* Paper presented at the ESOMAR Worldwide Internet Conference Net Effects 3, Dublin, Ireland.

[10] Converse, P. D., Wolfe, E. W., Xiaoting, H., and Oswald, F. L. (2008): Response rates for mixed-mode surveys using mail and e-mail/web. *American Journal of Evaluation*, **29**, 99-107.

[11] Cook, C., Heath, F., and Thompson, R. (2000): A meta-analysis of response rates in web- or Internet-based surveys. *Educational & Psychological Measurement,* **60**, 821-836.

[12] Couper, M. P., Kapteyn, A., Schonlau, M., and Winter, J. (2007): Noncoverage and nonresponse in an Internet survey. *Social Science Research,* **36**, 131-148.

[13] de Leeuw, E. D. (2005): To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* **21**, 233–255.

[14] de Leeuw, E. D. and de Heer, W. (2002): Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, Dillman, D. A., Eltinge, J. L. and Little, R. J. A. (Eds.): *Survey Nonresponse*, 41-54. New York, NY: John Wiley & Sons.

[15] Dever, J. A., Rafferty, A., and Valliant, R. (2008): Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods,* **2**, 47-62.

[16] Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014): *Internet, phone, mail and mixed-mode surveys: The tailored design method*. Hoboken, NJ: Wiley.

[17] Dolničar, S., Laesser, C., and Matus, K. (2009): Online versus paper: Format effects in tourism surveys. *Journal of Travel Research,* **47**, 295-316.

[18] Downes-Le Guin, T., P. Janowitz, R. Stone, and S. Khorram. (2002): Use of pre-incentives in an Internet survey. *Journal of Online Research*, **25**, 1-7.

[19] ESOMAR. (2014): *Global market research 2014*. Amsterdam: ESOMAR.

[20] Eurostat. (2015): *Individuals – internet use in the last three months* (Database). Luxembourg, LUX: Eurostat.

[21] Fricker, S. S., Galesic, M., Tourangeau, R., and Yan, T. (2005): An experimental comparison of Web and telephone surveys. *Public Opinion Quarterly*, **69**, 370-92.

[22] Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. (2008): Eye-tracking data. New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, **72**, 892-913.

[23] Göritz, A. (2006): Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science,* **1**, 58-70.

[24] Göritz, A. (2008): The long-term effect of material incentives on participation in online panels. *Field Methods,* **20**, 211-225.

[25] Göritz, A., Wolff, H.-G., and Goldstein, D. G. (2008): Individual payments as a longer-term incentive in online panels. *Behavior Research Methods,* **40**, 1144-1149.

[26] Greene, J., Speizer, H., and Wiitala, W. (2008): Telephone and web: Mixed-mode challenge. *Health Services Research,* **43**, 230-248.

[27] Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004): *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

[28] Groves, R. M. and Peytcheva, E. (2008): The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly,* **72**, 167-189.

[29] Heerwegh, D. and Loosveldt, G. (2008): Face-to-face versus web surveying in a high-Internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, **72**, 836-846.

[30] Internet World Stats (2014): *Internet users in the world.* Retrieved *30 June 2014*, from http://www.internetworldstats.com/stats.htm

[31] James, J. M. and Bolstein, R. (1992): Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly,* **56**, 442-453.

[32] Kish, L. (1965): *Survey sampling*. New York, NY: John Wiley & Sons.

[33] Knapp, H. and Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: does methodology matter? *Computers in Human Behavior*, *19*, 117–134. http://doi.org/10.1016/S0747-5632(02)00008-0

[34] Kreuter, F., Presser, S., and Tourangeau, R. (2008): Social desirability bias in CATI, IVR, and web surveys. The effects of mode and question sensitivity. *Public Opinion Quarterly*, **72**, 847-865.

[35] Kwak, N. and Radler, B. (2002). A comparison between mail and Web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics*, *18*, 257–73.

[36] Lee, S. and Valliant, R. (2009): Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research,* **37**, 319-343.

[37] Loosveldt, G. and Sonck, N. (2008): An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, **2**, 93-105.

[38] Lozar Manfreda, K. and Vehovar, V. (2002): Survey design features influencing response rates in web surveys. Paper presented at the ICIS 2002 The International Conference on Improving Surveys, Copenhagen, Denmark.

[39] Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008): Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research,* **50**, 79-104.

[40] Malhotra, N. (2008): Completion time and response order effects in web surveys. *Public Opinion Quarterly*, **72**, 914-934.

[41] McMorris, B. J., Petrie, R. S., Catalano, R. F., Fleming, C. B., Haggerty, K. P., and Abbott, R. D. (2009): Use of web and in-person survey modes to gather data from young adults on sex and drug use. *Evaluation Review,* **33**, 138-158.

[42] Medway, R. L. and Fulton, J. (2012): When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, **76**, 733-746.

[43] Messer, B. L. and Dillman, D. A. (2011): Surveying the General Public over the Internet Using Address-Based Sampling and Mail Contact Procedures. *Public Opinion Quarterly*, **75**, 429-457.

[44] Parackal, M. (2003): Internet-based & mail survey: A hybrid probabilistic survey approach. Paper presented at the AusWeb 2003: The 9th Australian World Wide Web Conference, Goald Coast, Australia.

[45] Parsons, N. L. and Manierre, M. J. (2014): Investigating the relationship among prepaid token incentives, response rates, and nonresponse bias in a web survey. *Field Methods*, **26**, 191-204.

[46] Porter, S. R. and Whitcomb, M. E. (2007): Mixed-mode contacts in web surveys: Paper is not necessarily better. *Public Opinion Quarterly*, **71**, 635-648.

[47] Rookey, B. D., Hanway, S., and Dillman, D. A. (2008): Does a probability-based household panel benefit from assignment to postal response as an alternative to Internet-only? *Public Opinion Quarterly*, **72**, 962-984.

[48] Sajti, A. and Enander, J. (1999): Online survey of online customers, value-added market research through data collection on the Internet. *Proceedings of the ESOMAR World-Wide Internet Conference Net Effects*, 35-51, Amsterdam: ESOMAR.

[49] Schonlau, M., van Soest, A., Kapteyn, A., and Couper, M. (2009): Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, **37**, 291-318.

[50] Shih, T.-H. and Fan, X. (2008): Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, **20,** 249-271.

[51] Singer, E. and Ye, C. (2013): The use and effects of incentives in surveys. The *ANNALS of the American Academy of Political and Social Sciences*, 645, 112-141.

[52] Smyth, J. D., Christian, L. M., and Dillman, D. A. (2008): Does "Yes or No" on the telephone mean the same as "Check-All-That-Apply" on the Web? *Public Opinion Quarterly*, **72**, 103-13.

[53] Tourangeau, R. (2004): Experimental design considerations for testing and evaluating questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.): *Methods for Testing and Evaluating Survey Questionnaires*, 209-224. Hoboken, New Jersey: John Wiley & Sons, Inc.

[54] Tourangeau, R., Couper, M. P., and Conrad, F. C. (2013): "Up means good": The effect of screen position on evaluative ratings in web surveys. *Public Opinion Quarterly*, **77**, 69-88.

[55] Vehovar, V., and Lozar Manfreda, K. (2008): Overview: Online surveys. In N. G. Fielding, R. M. Lee & G. Blank (Eds.): *The handbook of online research methods*, 177-194. Thousand Oaks, CA: SAGE Publications.

[56] Vehovar, V., Berzelak, N., and Lozar Manfreda, K. (2010): Mobile Phones in an Environment of Competing Survey Modes: Applying Metric for Evaluation of Costs and Errors. *Social Science Computer Review*, **28**, 303-318.

[57] Vehovar, V., Lozar Manfreda, K., and Batagelj, Z. (2001): Sensitivity of electronic commerce measurement to the survey instrument. *International Journal of Electronic Commerce*, **6**, 31-52.

[58] Villar, A. (2013): Feasibility of using web to survey at a sample of addresses: a UK ESS experiment. Paper presented at the Genpoweb, London, UK. Retrieved from http://www.natcenweb.co.uk/genpopweb/documents/ESS-experiment-Ana-Villar.ppt

Metodološki zvezki, Vol. 12, No. 2, 2015, 69-84

# A Review of Capital Structure Theory Using a Bibliometric Analysis

Denis Marinšek[1]

**Abstract**

Author citation and co-citation analysis is a simple, yet powerful educational tool for detecting the most relevant papers from any research field. I demonstrate its use and graphically show a chronological development of capital structure theory, which highlights the most important contributions. I then systematically present the capital structure theory, starting with Modigliani & Miller's irrelevance theorem, and continue with four alternative explanations of firm capital structure behaviour. This paper is particularly useful for PhD students and junior researchers who need to familiarize with the literature of their own research field, and for those, interested in the up-to-date review on capital structure theory.

## 1 Introduction

In a plethora of literature, finding key works and establish clear connections among them can be challenging. This paper addresses this issue, explaining how bibliometrics (i.e. citation and co-citation analysis with software BibExcel, and graphical presentation with software Pajek) can be applied to any research topic. The high importance of using innovative educational methods as learning aids was elaborated by Plumb & Zamfir (2011) and Pocatilu & Ciurea (2011). However, Dospinescu et al. (2011) found that students are often not enough informed about modern internet learning methods. I show that bibliometric analysis is a powerful educational tool, which offers a unique insight into a literature review.

Bibliometrics is used for detecting connections among different schools of thought and offers greater objectivity, which is a result of the outcome of a composite judgment of many citing authors (Bayer, Smart, & McLaughlin, 1990). Moreover, it helps to determine the most influential papers, detect leading scholars, and offer different graphical presentations of development of any research field (e.g. chronological overview, detecting theory streams, etc.). Besides, it can save a lot of time because it immediately directs a researcher to the most crucial papers, which should be the base for building a new research.

[1] Denis Marinšek, Faculty of Economics, Kardeljeva ploscad 17, 1000 Ljubljana; denis.marinsek@ef.uni-lj.si

In this paper bibliometrics is applied to the literature on capital structure theory. To the best of my knowledge this kind of analysis has never been performed on the capital structure literature. The graphical presentation of bibliometric findings shows that a modern capital structure theory began in the late fifties with the irrelevance theorem by Modigliani & Miller (1958), and that the theory developed from a neoclassical theory of the firm, which can be traced back into 1930s. As an answer to the irrelevance theorem, two theories of capital structure emerged: the trade-off theory (Jensen & Meckling, 1976), and the pecking order hypothesis (Myers & Majluf, 1984), both trying to explain observed behavior of firm's capital structure choices. More recently, the dynamic version of trade-off theory and equity market timing theory received a strong empirical support.

This paper has thus two goals. In section 2 it directs a reader to the literature on bibliometric tools, and demonstrates how they can be applied to facilitate a proper understanding of a chosen research topic. In section 3 it offers a comprehensive overview of capital structure theory from the beginnings to the most recently published articles. Conclusion summarizes the main findings of this paper.

# 2 Bibliometric analysis as an educational tool

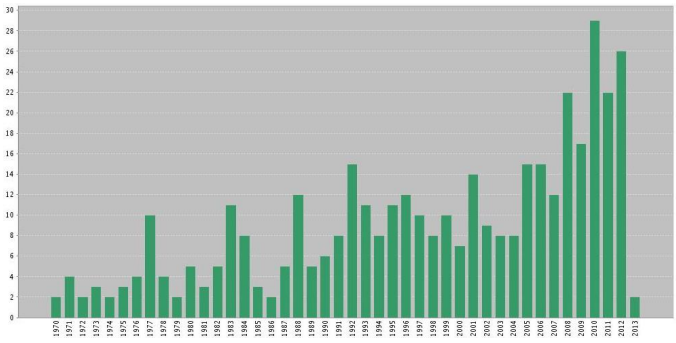
Citation and co-citation analysis is performed on 400 English papers from the ISI Web of Science database (Thomson Reuters, 2013), related to the capital structure theory and published until February 2013. The database consists of three citation sources: Science Citation Index Expended (1970–present), Social Sciences Citation Index (1970–present), and Arts & Humanities Citation Index (1975–present). At the time of downloading the database of articles, the last refresh of the database was on 22nd February 2013. The articles were found with the use of keywords: "Theory of capital structure" or "Modigliani-Miller theorem" or "Pecking order theory" or "Trade-off theory" or "Optimal debt level" or "Optimal leverage" or "Leverage and firm's performance" or "Financing decision" or "Target capital structure" or "Modern capital structure theory". With this search inquiry 8,980 articles were found. Out of all results, only English articles from four Web of Science categories were retained: Economics, Business Finance, Management, and Business. This step reduced the database to 4,120 articles. Further, abstracts of all potentially interesting articles were analyzed and articles that were not related to the capital structure theory were excluded. Finally, 400 most relevant articles for capital structure research were kept, which were the base for citation and co-citation analysis. For a detailed overview on various bibliometric methods, see White & Belvar (1981), Bayer, Smart & McLaughlin (1990) and De Bellis (2009).

## 2.1 Citation analysis

Figure 1 shows distribution of 400 primary articles by year of publishing (number of published articles in the year 2013 is not directly comparable since only

the first two months are included), showing that the majority of papers were published more recently. Over the observed period, the average number of citations per paper is 132.39 and h-index is 90. This means that there are 90 papers among 400 primary papers that have at least 90 citations

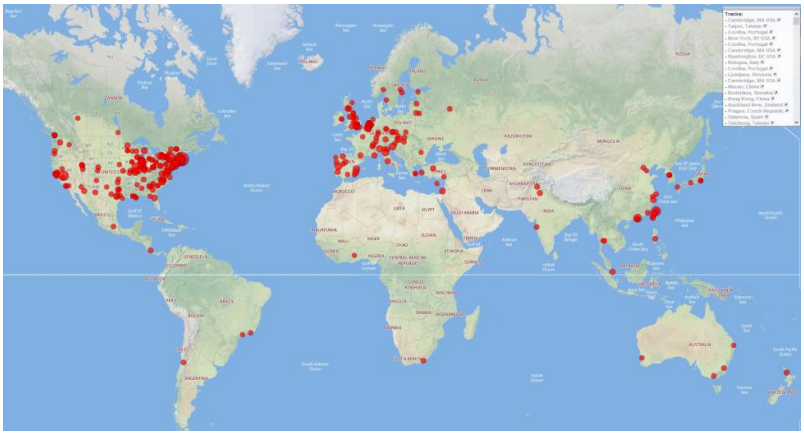
**Figure 1:** Distribution of primary papers by the year of publishing



*ISI Web of Science, 2013.*

Moreover, Figure 2 demonstrates where these 400 primary papers were published. This can be performed by extracting authors' addresses with software BibExcel, and then utilizing GPS Visualizer (Persson, 2009).

**Figure 2:** Geographical locations of authors of primary papers



*ISI Web of Science, 2013.*

## 2.2 Co-citation analysis

After analyzing primary papers, the analysis of the target papers (i.e. papers that are citied within primary papers) is performed - the co-citation analysis. This is the analysis of joint occurrence of target papers within primary papers. Co-occurrence analysis, performed with the use of BibExcel software, is therefore the study of mutual appearances of pairs of units in analyzed bibliographic records (Persson, Danell, & Wiborg Schneider, 2009). Authors, which co-occur together, are the base for the analysis of different schools of thoughts.

Table 1 shows the most important journals where target papers were published (i.e. frequency of occurrence of journals within the references of 400 primary

papers). This can help as an orientation where further research on this topic can be published.

**Table 1:** Journals, which published the target articles

| *frequency* | Journal name |
|---|---|
| 2723 | The Journal of Finance |
| 2005 | Journal of Financial Economics |
| 746 | American Economic Review |
| 369 | Review of Financial Studies |
| 350 | Journal of Political Economy |
| 271 | Journal of Financial and Quant. analysis |
| 236 | Financial Management |
| 211 | Quarterly Journal of Economics |
| 210 | Journal of Business |
| 195 | Bell Journal of Economics |
| 184 | Econometrica |
| 131 | Review of Economic Studies |

*ISI Web of Science, 2013.*

Additionally, Table 2 shows the frequency of citations of the most cited target articles by 400 primary articles, denoted by *f*. This is the list of the most influential papers in the capital structure theory.

**Table 2:** Most frequently cited references by 400 primary articles

| *f* | First author, year and publication | *f* | First author, year and publication |
|---|---|---|---|
| 169 | Jensen M, 1976, V3, P305, J Financ Econ | 49 | Stulz R, 1990, V26, P3, J Financ Econ |
| 144 | Modigliani F, 1958, V48, P261, Am Econ Rev | 48 | MacKie-Mason J, 1990, V45, P1471, J Financ |
| 141 | Myers S, 1977, V5, P147, J Financ Econ | 47 | Frank M, 2003, V67, P217, J Financ Econ |
| 136 | Myers S, 1984, V13, P187, J Financ Econ | 47 | Marsh P, 1982, V37, P121, J Financ |
| 122 | Titman S, 1988, V43, P1, J Financ | 46 | Baker M, 2002, V57, P1, J Financ |
| 117 | Rajan R, 1995, V50, P1421, J Financ | 44 | Fischer E, 1989, V44, P19, J Financ |
| 113 | Jensen M, 1986, V76, P323, Am Econ Rev | 42 | Graham J, 2001, V60, P187, J Financ Econ |
| 95 | Myers S, 1984, V39, P575, J Financ | 41 | Titman S, 1984, V13, P137, J Financ Econ |
| 85 | Harris M, 1991, V46, P297, J Financ | 36 | Leary M, 2005, V60, P2575, J Financ |
| 81 | Modigliani F, 1963, V53, P433, Am Econ Rev | 35 | Welch I, 2004, V112, P106, J Polit Econ |
| 80 | Bradley M, 1984, V39, P857, J Financ | 35 | Booth L, 2001, V56, P87, J Financ |
| 73 | Miller M, 1977, V32, P261, J Financ | 34 | Kraus A, 1973, V28, P911, J Financ |
| 73 | DeAngelo H, 1980, V8, P3, J Financ Econ | 32 | Smith C, 1979, V7, P117, J Financ Econ |
| 70 | Shyam-Sunder L, 1999, V51, P219, J Financ Econ | 32 | Graham J, 2000, V55, P1901, J Financ |
| 70 | Fama E, 2002, V15, P1, Rev Financ Stud | 31 | Flannery M, 2006, V79, P469, J Financ Econ |
| 67 | Ross S, 1977, V8, P23, Bell J Econ | 30 | Leland H, 1977, V32, P371, J Financ |
| 55 | Hovakim. A, 2001, V36, P1, J Financ Quant Anal | 30 | Jalilvand A, 1984, V39, P127, J Financ |

With information gathered from citation and co-citation analysis, and with the help of Pajek software (Batagelj & Mrvar, 1998; Persson, 2009b), Figure 3 presents 81 most important papers from the capital structure theory, i.e. papers that were most often cited by primary papers (data for drawing this figure are published as a supplementary information). The size of the circle represents the importance of a paper by number of citations, while the thickness of lines among papers depends on the number of co-citations among them. Additionally, papers are presented in a chronological order from the earliest papers on the top of the figure to the most recent

ones at the bottom (each year has a unique color). In the following sections of this paper, the theory of capital structure is presented by closely following the findings from Figure 3.

# 3 Modern capital structure theory

Capital structure theory emerged from the neoclassical theory of the firm, which began with Berle & Means (1932) and was continued by Coase (1937). On the basis of their work, two separate theories developed, both trying to explain what a firm is and how it operates. The first one is the agency cost theory (Jensen & Meckling, 1976) that advocates a firm's existence because of its positive effects, created by a team production. The second one is the property rights theory (Demsetz, 1967) that concentrates on the contractual relationships within the firm. Both theories were the foundation for the modern capital structure theory, which began in the late fifties with the irrelevance theorem.

In 1958, Modigliani & Miller (hereafter M&M) published an influential article The cost of capital, corporation finance and the theory of investments, which was based on the neoclassical definition of the firm. Before this work, there was no generally accepted theory of the capital structure (Frank & Goyal, 2008). In the article, authors assumed numerous unrealistic assumptions, which were, however, gradually omitted in their further publishing (e.g. maximization of the shareholders' value is the only goal of a firm; firm is financed only with equity and debt; there are no taxes; individuals can borrow and lend at the risk-free rate; markets have no frictions; firms operate in competitive markets; there is no asymmetric information; etc.).

Under these assumptions two propositions were made. The first proposition says that the value of a levered firm is equal to the value of an unlevered firm. M&M argued that if two firms were identical (if generated the same cash flow), but differed only in their capital structures, then arbitrage opportunities would force values of both firms to become equal. As a result, the leverage has no effect on the market value of a firm. However, Baxter (1967) soon argued that with a high level of indebtedness, the 'risk of ruin' becomes very real and cannot be nullified by arbitrage. He concluded that when reliance on the financial debt is small, the tax-shelter effect dominates, but as soon as leverage increases too much, risk of ruin prevails. The same conclusion was proposed by Donaldson (1961), Robichek & Myers (1966) and Kraus & Litzenberger (1973). The second M&M's proposition says that benefits, obtained from the increased use of the low cost debt and decreased use of the high cost equity, are completely offset by the increase in the risk level of equity. Shortly, leverage has no effect on the cost of capital.

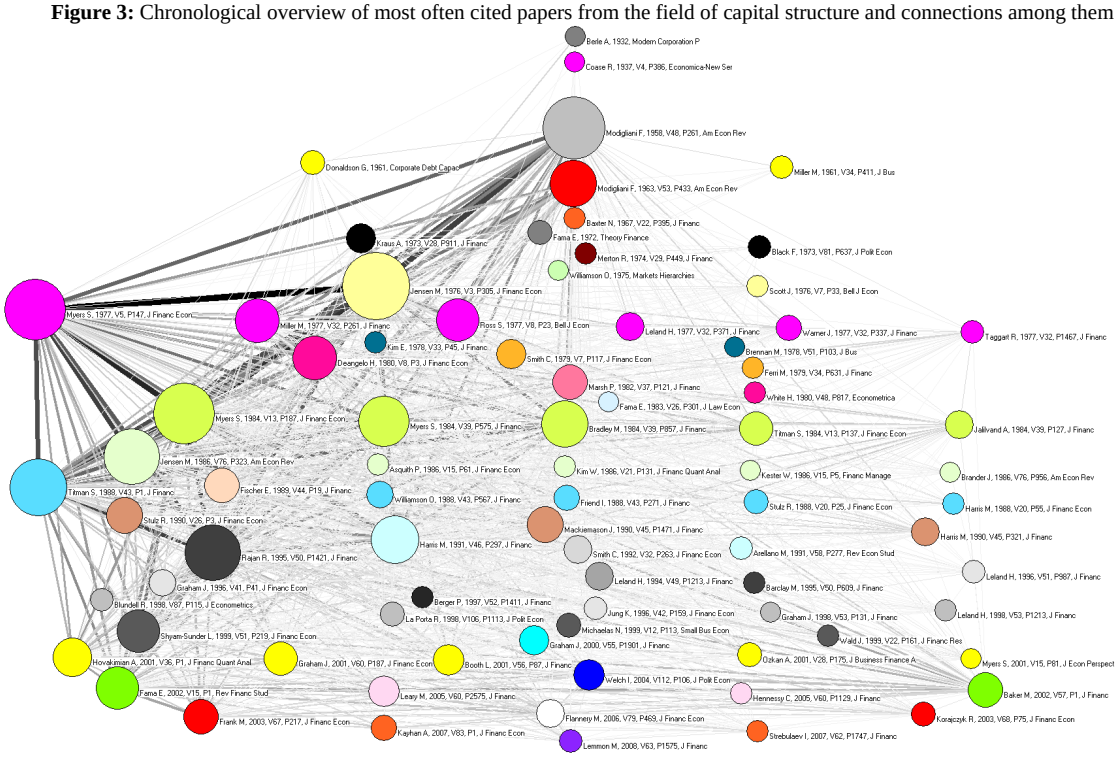
In 1961, M&M developed the dividend irrelevance proposition, which implied that the value of a firm is unaffected by the distribution of dividends, but is solely determined by the earning power and risk of its assets (Miller & Modigliani, 1961). Two years later, M&M (1963) developed the investor indifference proposition, which says that equity-holders are indifferent about a firm's financial policy, the

thesis which was later deeply elaborated by Stiglitz (1969; 1974). In their latest work, M&M introduced corporate taxes and showed that leverage increases a firm's value because interest costs are tax-deductible and consequently increase the income available to the shareholders. M&M were, however, careful about implying that the value of the firm would be maximized when using 100 percent debt financing. They argued that some other forms of financing, like retained earnings, can be cheaper than debt, and that lenders can impose limitations that prevent too high indebtedness. They concluded that firms need to preserve a certain rate of flexibility in maintaining reserve borrowing capacity (Modigliani & Miller, 1963).

Miller (1977) later introduced the effect of personal taxes and argued that in equilibrium, tax advantage of debt would be exactly offset by the increased personal taxation, which means that a shareholder would be indifferent to how much leverage the firm uses. He argued that if the optimal capital structure is simply a matter of rebalancing tax advantages against bankruptcy costs, why then the observed capital structures show so little variation over the time. Contrary, DeAngelo & Masulis (1980) argued that Miller's theorem is extremely sensitive to the realistic and simple modifications in the corporate tax code. They included into the analysis a tax shield that is not a result of the interest costs (e.g. accounting depreciation, depletion allowance, and investment tax credits) and showed that there is a market equilibrium, where every firm has a unique optimal capital structure. DeAngelo & Masulis continued that market prices reflect personal and corporate taxes in such a way that the bankruptcy costs are a significant consideration in a trade-off between interest tax-deductibility and risk of financial distress.

Modigliani (1980) summarized the M&M's irrelevance theorem in the following way: "… with well-functioning markets (and neutral taxes) and rational investors, who can 'undo' the corporate financial structure by holding positive or negative amounts of debt, the market value of the firm – debt plus equity – depends only on the income stream generated by its assets. It follows, in particular, that the value of the firm should not be affected by the share of debt in its financial structure or by what will be done with the returns – paid out as dividends or reinvested (profitably)". However, at the same time Chen & Kim (1979) made a synthesis of theoretical and empirical research, and figured out that the theory somehow acknowledges the benefits of debt on aggregate level but is unable to answer why firms are using risky debt on the individual level. Therefore it soon became clear that M&M's irrelevance theorem could not exist in a real economy, and researchers came to the conclusion that the capital structure must be relevant for a market value of a firm. Different theories emerged, explaining which factors are the most relevant when management is trying to find the optimal source of financing, i.e. the capital structure that would maximize the market value of a firm. The two most important theories are the trade-off theory and the pecking order hypothesis.

**Figure 3:** Chronological overview of most often cited papers from the field of capital structure and connections among them




*ISI Web of Science, 2013 & BibExcel & Pajek.*

## 3.1 Trade-off theory

Donaldson (1961) wrote an influential book called Corporate Debt Capacity, where he acknowledged that setting an appropriate limit for the borrowed amount of long-term debt is a major challenge for financial management. The reason is in the importance that a debt-equity ratio has for future solvency and profitability of a firm. Donaldson therefore tried to address a problem firms face every day: "Given the need for new permanent capital and ability to borrow, how does a company approach the determination of the wise and proper limit to such borrowing?" He argued that the main determinant of corporate debt capacity should be the probability of insolvency in times of recession, analyzed thorough firm's cash flows. His reasoning is popular even today, as for example in Kester et al. (2004), who argued that debt capacity should not be determined solely by industry averages or the availability of collateral, but also by borrower's ability to repay interests and the principal with a generated cash flow. Donaldson introduced the concept of risk and fear of insolvency with the help of objective risk assessment (management of cash flows) and subjective risk assessment (management inclination to risk-taking). It is generally accepted that the primary incentive to use long-term debt is the fact that debt is theoretically a cheaper source of financing than retained earnings or new equity issues. If the primary objective of a business is the maximization of net revenues, Donaldson (1961) continued, the use of debt should be a desirable source of financing and should be exercised as continuously as possible. The advantage of debt financing was especially well recognized in the period of high taxes during and after the World War 2 (Donaldson, 1961). It can be concluded that Donaldson was one of the first researchers who argued that the capital structure must be determined by weighing positive and negative effects of debt financing.

The trade-off theory developed as a response to the M&M's irrelevance theorem. When M&M (1963) added the effect of corporate tax, while ignoring the offsetting costs of debt (increased possibility of financial distress), 100 percent leverage was suggested as an optimal capital structure, although this was in contradiction with the observed firms' behavior. Kraus & Litzenberg (1973) suggested that the best candidate for an offsetting cost of debt is a deadweight cost of bankruptcy, while Jensen & Meckling (1976) more formally defined two types of conflicts, one resulting as a debt benefits and other as costs of debt financing. The first conflict, highly elaborated by Donaldson (1963), is between shareholders and managers (principal-agent problem), and is expressed in a form of debt benefit. The reasoning is that higher leverage reduces the principal-agent problem, because managers have less available free cash flow to invest it unwisely. The second conflict, Jensen & Meckling continued, is between debt-holders and equity-holders, thoroughly presented by Smith & Warner (1979). In that case, higher leverage increases agency costs of debt, because benefits are borne primarily by equity-holders, while costs by debt-holders. Optimal capital structure can then be found by trading-off benefits and costs of debt, which is the basic idea of the static version of trade-off theory. Myers (1984) later argued that a firm, which follows the trade-off theory, sets target

leverage and then gradually moves toward that target – the main idea of the dynamic version of trade-off theory, which has become popular more recently.

Jensen & Meckling (1976) claimed that M&M (1963) were unable to offer an adequate theory of the observed capital structures. Similarly, Fama & Miller (1972) wrote the following sentence: "We must admit that at this point there is little in the way of convincing research, either theoretical or empirical, that explains the amount of debt that firms do decide to have in their capital structure." Jensen & Meckling (1976) argued that agency costs provide strong reasons for dependency between probability distribution of future cash flow and capital/ownership structure. They argued that while introduction of bankruptcy costs and presence of tax subsidies leads to the theory of optimal capital structure, that theory has serious drawback since it implies that no debt should ever be used in the absence of tax subsidies in case of positive bankruptcy costs. However, because firms have been using debt already when there were no tax benefits on interest costs, there must be some additional determinants of corporate capital structure. Moreover, neither the bankruptcy costs nor the tax subsidies can explain the use of preferred stocks or warrants, even more, theory says nothing about division of equity claims between insiders and outsiders. Researchers thus started arguing that bankruptcy costs themselves are unlikely to be the main determinant of a firm's capital structure because empirical research showed (e.g. Warner (1977)) that these costs represent very small percent of a firm's value. On the other hand, Baxter (1967) and Stanley & Girth (1971) showed that for smaller firms this percentage can be considerably higher. Furthermore, Kim (1978) was one of the important advocators of the idea that firm reaches an optimal capital structure at level of indebtedness far below theoretically proposed 100 percent, as argued by some researchers before him. He continued that only when the target debt level is strictly lower than a firm's debt capacity, the firm can search for its optimal trade-off between the tax advantage of debt and the cost of bankruptcy. Value of a firm can therefore be defined as the value of equity only financed firm, increased for the value of the tax savings and decreased for the value of the costs of financial distress. Value of the firm is therefore maximized at less than 100 percent debt financing (Morris, 1982).

According to the trade-off theory, there exists the optimal capital structure. The theory describes the firm's optimal capital structure as a mix of financing that equates the marginal costs to marginal benefits of debt financing (Lemmon & Zender, 2010). However, it is important to distinguish between the static trade-off theory, where firm balances tax savings of debt against deadweight bankruptcy costs, and more recently introduced dynamic trade-off theory. A firm is said to follow the static trade-off theory when firm's leverage is determined by a single period trade-off between the tax benefits of debt and the deadweight costs of bankruptcy (Frank & Goyal, 2008). However, the main drawback of this theory is that it says nothing about a mean reversion of leverage. Consequently, the dynamic trade-off theory developed, proposing that a firm exhibits the target adjustment behavior if it has the target leverage and if deviations from that target are gradually eliminated over the time. The dynamic trade-off theory has advanced in recent years because it offers a good

explanation of tendency movements of leverage, the role of profits, the role of retained earnings and the path dependency. Frank & Goyal (2008) concluded that the target adjustment hypothesis receives much better empirical support than did either the static trade-off theory or the pecking order hypothesis, which is presented in the next section. An interesting overview of convergence toward the target capital structure is given by Lemmon et al. (2008). They clearly show that although leverage ratios exhibit persistence, there is a strong convergence toward the moderate indebtedness. Moreover, many researchers numerically estimated the speed of convergence. The majority of past empirical research on convergence is based on partial adjustment models, where average yearly rate of adjustment toward the predefined target is estimated, i.e. the speed. The most recent published speed estimates are 31 percent per year by Flannery & Rangan (2006), 25 percent by Lemmon et al. (2008), 23 percent by Huang & Ritter (2009), 22 percent by Byoun (2008), 7-18 percent by Fama & French (2002), to practically zero by Baker & Wurgler (2002), and Welch (2004). Nevertheless, Frank & Goyal (2009) performed a comprehensive review of past empirical studies that examined the determinants that had a significant power at explaining observed capital structures and that gave consistent findings over many tests (e.g. Rajan & Zingales (1995)). The six main determinants are industry median leverage (firms in industries in which the median firm has high leverage tend to have higher leverage), tangibility (firms that have more tangible assets tend to have higher leverage), profits (firms that have more profits tend to have lower leverage), firm size (firms that have larger assets or higher sales tend to have higher leverage), market-to-book-assets ratio (firms that have a high market to book ratio tend to have lower leverage), and inflation (when inflation is expected to be high, firms tend to have high leverage). Frank & Goyal concluded that these six determinants explain 27 percent of the variation of leverage.

## 3.2 Pecking order hypothesis

Conflicts, related to the existence of inside information, were the main driver of the development of the theory of asymmetric information, under which managers and owners have more accurate information about a firm's true performance than those who lend the money. This theory has evolved into two directions. The first direction is connected to Ross (1977) and Leland & Pyle (1977), who argued that the choice of a firm's capital structure gives an important signal (inside information) to outside investors. The second direction is represented by Myers & Majluf (1984) and Myers (1984), who argued that the capital structure is used to solve inefficiencies in the firm's investment decisions, which are caused by the information asymmetry. This direction is associated with the so called pecking order hypothesis (Myers & Majluf (1984), Krasker (1986) and Narayanan (1988)). The costs and benefits of debt (trade-off theory) are of secondary importance compared to the importance of costs which arise when a new equity is issued under the conditions of highly asymmetric information (Shyam-Sunder & Myers, 1999).

Myers & Majluf (1984) advocated the idea that if the potential new shareholder is

not equally informed as the existing shareholders, the former one will underprice new equity issues, the problem called "adverse selection". That would primarily cause a loss to the existing shareholders. The problem can be mitigated by giving priority to all other types of financing before issuing a new equity. These other types of financing are retained earnings and different forms of debt. This behavior is called the pecking order hypothesis. As a direct consequence, share price should fall after the announcement of new equity issue. Krasker (1986) confirmed this finding and additionally showed that the larger the stock issue, the larger will the fall in stock price be. This problem can result in underinvestment, which is more severe for firms with relatively low levels of tangible assets (Harris & Raviv, 1991). Myers (1984) summarized the pecking order hypothesis: a firm is said to follow a pecking order if it prefers internal to external financing and debt to equity, when external financing is used. A more recent empirical analysis of the pecking order theory was performed by Frank & Goyal (2003), who found that internal financing is often not sufficient to cover investment spending, which means that external financing is heavily used, often prioritizing debt.

## 3.3 Industrial organization

Capital structure models that are based on industrial organization theory can be divided into two groups. The first group of research explains relations between firm's capital structure and firm's strategy, while the second group of research explains relations between firm's capital structure and the characteristics of its products and inputs (Harris & Raviv, 1991). Brander & Lewis (1986) and Maksimovic (1988) were the initiators of the idea that financial theory of maximizing shareholders' value can be linked to industrial organization, where researchers typically used assumption of maximization of a total profit. These authors referred to the finding of Jensen & Meckling (1976) that increased leverage encourages equity holders to accept riskier strategies. In Brander & Lewis (1986) model, oligopolists increased the risk through aggressive production policy, and in order to finance it, firms in a subsequent Cournot game choose higher and higher levels of debt. As a result, Brander & Lewis argued, oligopolists will often use more debt financing compared to monopolists or firms in competitive markets. Additionally, debt will be of long-term nature. Maksimovic (1988) also proved that debt capacity is increased with the elasticity of the demand.

The second group of research is concentrated around Titman's (1984) observations that customers and suppliers of unique and durable products would bare higher costs if a firm goes bankrupt, which means that such firms will be less indebted, ceteris paribus. When it is likely that a firm will stop operating, these costs are transferred to the shareholders through lower product prices. Titman argued that capital structure can be used to commit the shareholders to have an optimal liquidation policy. He showed that firms with higher costs of liquidation will use lower amounts of debt. Maksimovic & Titman (1991) found evidence that even consumer of non-durable goods and services (e.g. hospitals, pharmaceuticals, and air travels) are concerned

with the financial status of the producer, especially because of safety issues (e.g. in order to avoid bankruptcy, firm can reduce the quality of the products).

### 3.4 Market timing theory

The market timing plays an important role in describing observed capital structures. That was proposed already by Myers (1984), but it became more popular recently (e.g. Berry et al. (2008)). Graham & Harvey (2001) found empirical support that management actively uses market timing when deciding whether to issue debt or equity – they found that firms issue equity after the increases of stock prices. Baker & Wurgler (2002) argued that capital structure can best be understood as the cumulative effect of past attempts to time the market. Frank & Goyal (2009) summarized this theory as management analyzing the current market conditions in debt and equity markets. When a firm needs new financing, management uses the type of financing which is more favorable at the moment. If neither of them looks favorable, management can defer the issuances. On the other hand, if current conditions look unusually favorable, funds may be raised even if the firm currently does not need new funds. The shortfall of this theory is that it cannot be linked with the traditional determinants of capital structure; however, it suggests that stock returns and debt market conditions are important when management is evaluating capital structure decisions (Frank & Goyal, 2009).

## 4 Conclusions

This paper demonstrates how bibliometrics can be used as an educational tool, and be applied to a chosen field of literature. Firstly, a comprehensive dataset of papers can be obtained from ISI Web of Science. Secondly, once a set of relevant papers is selected, software BibExcel offers a great variety of tools for performing bibliometric analyses. Thirdly, software Pajek allows numerous graphical presentations of bibliometric information.

The graphical analysis shows (see Figure 3) that the modern capital structure theory began in the late fifties with the irrelevance theorem, and that it emerged from the neoclassical theory of the firm, which can be traced back into 1930s with works of Berle & Means (1932) and Coase (1937). In 1958, Modigliani & Miller introduced the irrelevance theorem, which stated that a capital structure does not affect a firm's value. The theorem was later modified with the inclusion of tax-deductibility of interest on debt (i.e. tax shield), which led to the conclusion that a firm's value is maximized at 100 percent of debt financing. Since the theorem was in contradiction with observed behavior, many researchers started arguing that in the real world capital structure does matter. The main argument was that with a high level of leverage, the risk of financial distress becomes significant and real. As a result, two theories of a firm's capital structure emerged: the trade-off theory (Jensen & Meckling, 1976), and the pecking order hypothesis (Myers & Majluf, 1984), both

trying to explain a firm's observed behavior of capital structure choices. The basic idea of the trade-off theory is that firm's optimal capital structure is determined as the mix of financing that equates the marginal costs to marginal benefits of debt financing (Lemmon & Zender, 2010). More recently, the dynamic version of trade-off theory received high attention. It concentrates on gradual adjustments toward the target capital structure, and tries to explain temporary deviations from the target, defined by traditional trade-off determinants (e.g. tangibility of assets, profitability, size, etc.). A strong gradual convergence toward the target capital structure was empirically shown by Lemmon et al. (2008). Frank & Goyal (2008) concluded that the target adjustment hypothesis receives a strong empirical support. In parallel to the trade-off theory, the pecking order hypothesis was developed. It prescribes the order of financing, which would maximize a firm's value. Myers (1984) summarized that a firm is said to follow a pecking order if it prefers internal to external financing and debt to equity, when external financing is used. Unlike the trade-off theory, it assumes that a firm does not have a target capital structure. Furthermore, two alternative theories for explanation of observed capital structure behavior emerged. First theory concentrates on products a firm produces. Firms that produce unique, durable products were found to have statistically lower indebtedness, because such firms have high costs of financial distress. Similarly, firms that offer products and services where safety issues are of high importance (e.g. hospitals, pharmaceuticals, and air travels) are less indebted (Maksimovic & Titman, 1991). Second theory, equity market timing, was highly elaborated by Baker & Wurgler's (2002). They found that past variation in market-to-book ratio has the strongest explanatory power of observed capital structure. The idea is that management is trying to exploit irrational investors by issuing equity when they are overly enthusiastic (e.g. Graham & Harvey (2001)).

# References

[1] Baker, M., & Wurgler, J. (2002). Market Timing and Capital Structure. *The Journal of Finance,* **57** (1), 1-32.

[2] Batagelj, V., & Mrvar, A. (1998). Pajek - program for large network analysis. *Connections,* **21** (2), 47-57.

[3] Baxter, N. D. (1967). Leverage, Risk of Ruin and Cost of Capital. *The Journal of Finance,* **22** (3), 395-403.

[4] Bayer, A. E., Smart, J. C., & McLaughlin, G. W. (1990). Mapping Intellectual Structure of a Scientific Subfield through Author Cocitations. *Journal of the American Society for Information Science,* **41** (6), 444-452.

[5] Berle, A. A., & Means, G. C. (1932). *The Modern Corporation and Private Property.* New Jersey: Transaction Publishers.

[6] Berry, C. B., Mann, S. C., Mihov, V. T., & Rodriguez, M. (2008). Corporate debt issuance and the historical level of interest rates. *Financial Management,* **37** (3), 413-430.

[7] Brander, J. A., & Lewis, T. R. (1986). Oligopoly and Financial Structure: The Limited Liability Effect. *American Economic Review,* **76** (5), 956-970.
[8] Byoun, S. (2008). How and When do Firms Adjust their Capital Structures towards Targets? *The Journal of Finance,* **63** (6), 3069-3096.
[9] Chen, A. H., & Kim, H. E. (1979). Theories of Corporate Debt Policy: A Synthesis. *The Journal of Finance,* **34** (2), 371-384.
[10] Coase, R. H. (1937). The Nature of the Firm. *Economica,* **4** (16), 386-405.
[11] De Bellis, N. (2009). *Bibliometrics and Citation Analysis.* Lanham: Scarecrow Press.
[12] DeAngelo, H., & Masulis, R. W. (1980). Optimal Capital Structure under Corporate and Personal Taxation. *Journal of Financial Economics,* **8** (1), 3-29.
[13] Demsetz, H. (1967). Toward a Theory of Property Rights. *American Economic Review,* **57** (2), 347-359.
[14] Donaldson, G. (1961). *Corporate Debt Capacity: A study of corporate debt policy and the determination of corporate debt capacity.* Boston: Harvard University.
[15] Donaldson, G. (1963). Financial Goals: Management vs. Stockholders. *Harvard Business Review,* **41** (3), 116-129.
[16] Dospinescu, N., Tătăruşanu, M., Butnaru, G. I., & Berechet, L. (2011). The Perception of Students from the Economic Area on the New Learning Methods in the Knowledge Society. *The AMFITEATRU ECONOMIC journal,* **13** (30), 527-543.
[17] Fama, E. F., & French, K. (2002). Testing the trade-off and the Pecking order Predictions about Dividends and Debt. *Review of Financial Studies,* **15** (1), 1-33.
[18] Fama, E. F., & Miller, M. H. (1972). *The theory of finance.* New York: Holt, Rinehart and Winston.
[19] Flannery, M. J., & Rangan, K. P. (2006). Partial adjustments toward target capital structures. *Journal of Financial Economics,* **79** (3), 469-506.
[20] Frank, M. Z., & Goyal, V. K. (2003). Testing the pecking order theory of capital structure. *Journal of Financial Economics,* **67** (2), 217-248.
[21] Frank, M. Z., & Goyal, V. K. (2008). Trade-off and Pecking Order Theories of Debt. In E. Eckbo, *The Handbook of Empirical Corporate Finance* (pp. 135-197). Elsevir Science.
[22] Frank, M. Z., & Goyal, V. K. (2009). Capital Structure Decisions: Which Factors are Reliably Important? *Financial Management,* **38** (1), 1-37.
[23] Graham, J. R., & Harvey, C. R. (2001). The theory and practice of corporate finance: evidence from the field. *Journal of Financial Economics,* **60** (2-3), 187-243.
[24] Harris, M., & Raviv, A. (1991). The Theory of Capital Structure. *The Journal of Finance,* **46** (1), 297-355.
[25] Hunag, R., & Ritter, J. R. (2009). Testing Theories of Capital Structure and Estimating the Speed od Adjustment. *Journal of Financial and Quantitative Analysis,* **44** (2), 237-271.
[26] Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency costs and Ownership Structure. *Journal of Financial Economics,* **3** (4), 305-360.

[27] Kester, G. W., Hoover, S. A., & Pirkle, K. M. (2004). How Much Debt Can a Borrower Afford? *The RMA Journal,* **87** (3), 46-51.
[28] Kim, E. H. (1978). A Mean-Variance Theory of Optimal Capital Structure and Corporate Debt Capacity. *The Journal of Finance,* **33** (1), 45-63.
[29] Krasker, W. S. (1986). Stock Price Movements in Response to Stock Issues under Asymmetric Information. *The Journal of Finance,* **41** (1), 93-105.
[30] Kraus, A., & Litzenberger, R. H. (1973). A State-Preference Model of Optimal Financial Leverage. *The Journal of Finance,* **28** (4), 911-922.
[31] Leland, H. E., & Pyle, D. H. (1977). Informational Asymmetries, Financial Structure, and Financial Intermediation. *Journal of Finance,* **32** (2), 371-387.
[32] Lemmon, M. L., & Zender, J. F. (2010). Debt Capacity and Tests of Capital Structure Theories. *Journal of Financial and Quantitative Analysis,* **45** (5), 1161-1187.
[33] Lemmon, M. L., Roberts, M. R., & Zender, J. F. (2008). Back to the Beginning: Persistence and the Cross-Section of Corporate Capital Structure. *The Journal of Finance,* **63** (4), 1575-1608.
[34] Maksimovic, V. (1988). Capital Structure in Repeated Oligopolies. *RAND Journal of Economics,* **19** (3), 389-407.
[35] Maksimovic, V., & Titman, S. (1991). Financial Policy and reputation for product quality. *Review of Financial Studies,* **4** (1), 175-200.
[36] Miller, M. H. (1977). Debt and Taxes. *The Journal of Finance,* **32** (2), 261-275.
[37] Miller, M. H., & Modigliani, F. (1961). Dividend Polity, Growth, and the Valuation of Shares. *Journal of Business,* **34** (4), 411-433.
[38] Modigliani, F. (1980). Introduction. In A. Abel, *The Collected Papers of Franco Modigliani, volume 3* (pp. xi-xix). Cambridge, Massachusetts: MIT press.
[39] Modigliani, F., & Miller, M. (1958). The cost of Capital, Corporation Finance and the Theory of Investment. *The American Economic Review,* **48** (3), 261-297.
[40] Modigliani, F., & Miller, M. (1963). Corporate Income Taxes and the Cost of Capital: A correction. *The American Economic Review,* **53** (3), 433-443.
[41] Morris, J. R. (1982). Taxes, Bankruptcy Costs and the Existence of an Optimal Capital Structure. *The Journal of Financial Research,* **5** (3), 285-299.
[42] Myers, S. C. (1984). The Capital Structure Puzzle. *The Journal of Finance,* **39** (3), 575-592.
[43] Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investments decisions when firms have information the investors do not have. *Journal of Financial Economics,* **13** (2), 187-221.
[44] Narayanan, M. P. (1988). Debt Versus Equity under Asymmetric Information. *Journal of Financial and Quantitative Analysis,* **23** (1), 39-51.
[45] Persson, O. (2009). Making Google Maps. Retrieved 3 1, 2013, from http://www8.umu.se/inforsk/Bibexcel/
[46] Persson, O. (2009). Mapping science using Bibexcel and Pajek. European Summer School for Scientometric. Retrieved 2 28, 2013, from http://www8.umu.se/inforsk/Bibexcel/
[47] Persson, O., Danell, R., & Wiborg Schneider, J. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Astrom, R. Danell, B. Larsen,

& J. Wiborg Schneider, *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday* (pp. 9-24). ISSI.

[48] Plumb, I., & Zamfir, A. (2011). A Possible Model for Developing Students' Skills within the Knowledge-Based Economy. *The AMFITEATRU ECONOMIC journal,* **13** (30), 482-496.

[49] Pocatilu, P., & Ciurea, C. (2011). Modern Solutions For Economic Higher Education In The Knowledge-Based Society. *The AMFITEATRU ECONOMIC journal,* **13** (30), 497-511.

[50] Rajan, R. G., & Zingales, L. (1995). What Do We Know about Capital Structure? Some Evidence from International Data. *Journal of Finance,* **50** (5), 1421-1460.

[51] Robichek, A. A., & Myers, S. C. (1966). Problems in the Theory of Optimal Capital Structure. *Journal of Financial and Quantitative Analysis,* **1** (2), 1-35.

[52] Ross, S. A. (1977). The Determination of Financial Structure: The Incentive-Signalling Approach. *The Bell Journal of Economics,* **8** (1), 23-40.

[53] Shyam-Sunder, L., & Myers, S. C. (1999). Testing static tradeoff against pecking order models of capital structure. *Journal of Financial Economics,* **51** (2), 219-244.

[54] Smith, C. W., & Warner, J. B. (1979). An analysis of Bond Covenants. *Journal of Financial Economics,* **7**, 117-161.

[55] Stanley, D. T., & Girth, M. (1971). *Bankruptcy: Problem, Process, Reform.* Washington: Brookings Institution.

[56] Stiglitz, J. E. (1969). A Re-Examination of the Modigliani-Miller Theorem. *The American Economic Review,* **69** (5), 784-793.

[57] Stiglitz, J. E. (1974). On the Irrelevance of Corporate Financial Policy. *The American Economic Review,* **64** (6), 851-866.

[58] Thomson Reuters. (2013, 2 25). The Web of Science. New York, USA.

[59] Titman, S. (1984). The Effect of Capital Structure on a Firm's liquidation decision. *Journal of Financial Economics,* **13** (1), 137-151.

[60] Warner, J. B. (1977). Bankruptcy Costs: Some Evidence. *Journal of Finance,* **32** (2), 337-347.

[61] Welch, I. (2004). Capital Structure and Stock Return. *Journal of Political Economy,* **112** (1), 106-131.

[62] White, H. D., & Belver, G. C. (1981). Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science,* **32** (3), 163-171.

Metodološki zvezki, Vol. 12, No. 2, 2015, 85-110

# Design and Development of the Income Measures in the European Social Surveys

Juergen H.P. Hoffmeyer-Zlotnik[1] and Uwe Warner[2]

**Abstract**

In social surveys "total net household income" is an indicator of the respondent's socio-economic status. It describes the economic situation of household members and their positions in an income distribution. It is used as an explanatory variable in mobility studies and as a social-demographic background item in inequality research. This paper shows the impact of questionnaire design on measurements of "total net household income" in social surveys. In particular we illustrate how the measurement quality of the income variable depends on the data sources about the national income distributions used to design the answer categories offered to the respondent. Beginning with the fourth round of European Social Survey fielded in 2008 and the following years, the income categories for the question about the "total net household income" amount are built on national income distribution of households resident in the country under study. The response categories of the modified ESS questionnaires have been based on deciles of the actual household income distribution in the country in question. The central organizers of European Social Survey (ESS) instruct the national questionnaire designers to define the income brackets for the answer categories using the deciles of the most reliable national income data source. Analyzing the ESS data from 2008, 2010 and 2012, we found in some countries remarkable divergences from the expected 10% frequencies in each category. In this article we argue that the quality of this new income measure depends on the quality of the reference statistics from which the national household income ranges are derived. The quality of the responses to the survey question about the "total net household income", and finally the quality of the obtained survey measure, depends on the quality of the reference statistics from which the household income categories for the answers is derived. These reference data must cover all types of the household´s income from all household members and optimally represent the national distribution of household income across the survey universe. That means first that all possible payments accruing to a household and all its members in a given country must be reported in references, and second that all households

[1] Institute of Political Science, Justus Liebig University of Giessen, Germany, juergen.hoffmeyer-zlotnik@sowi.uni-giessen.de

[2] Perl, Germany, uwe,warner@orange.lu

in the survey's universe must be represented in the statistics used to detect the answer categories. Then the income brackets for the response categories can be calculated using the 10% percentiles from the income distribution in the reference data. Relevance, accuracy, timeliness, comparability, coherence, accessibility and clarity are quality domains of official statistics used as reference data for the survey measurement. We conclude that the central coordinators of the ESS define and communicate minimum threshold values for quality domains of the reference data. The national coordinators should report deviations. This would give the users of ESS data an insight into the quality of the income measurement.

# 1 Introduction

In a previous article (Hoffmeyer-Zlotnik and Warner, 2006) we discussed the measurement of "total net household income" in the first round of the European Social Survey (ESS) fielded in 2002 and we presented a proposal for changing the survey instrument in such a way that cross-national comparative social research with this socio-economic background variable would produce robust results.

One advantage of the questionnaire module on "total net household income" we proposed is that it includes a question that captures the impact of household size and the respondent's relationship to the main income earner. Moreover, the composition of the household income, the entire number of income types received by each member of the household in which the respondent lives, and the main source of income are collected using showcards. This helps the respondent to recall the individual components of the income received by all the household members and to add up these amounts to yield the "total net household income".

The second advantage of our proposal accounts for the various national income distributions of the participating countries. In countries with different income distributions and different average incomes, the income brackets of the response categories differ across the countries. Our main argument is that the lower income categories are more differentiated in countries with lower average income. In countries with higher average income crude income brackets can be used as answer categories at the lower part of the income distribution, however the categories at the upper part income distribution are differentiated. The ESS 2002 used one common showcard with the same answer categories and income brackets for all participating countries. To a certain extent, our suggestions for improvement were taken into account in the design of the fourth round of ESS implemented in 2008 and in subsequent rounds. The income categories for the question about income level are based on the national income distribution of households resident in the country in question. The response categories of the modified ESS questionnaires are based on deciles of the actual household income distribution in the country in question. In the present article we argue that the quality of this new income measure depends on the quality of the statistics from which the national household income ranges are derived.

## 2 Theoretical background: "Total net household income" as a socio-economic variable in social surveys

Demographic and socio-economic measures are so-called background variables. As Braun and Mohler (2003: 101f.) point out, "… they are used as independent variables, as socio-economic covariates of attitudes, behaviour, or test scores, etc. and in all sorts of statistical models, in particular, as endogenous factors in causal analysis. They enable analysts to establish homogeneous subgroups, explain differences of scale scores due to different composition, and to identify spurious correlations and causal relationships. Background variables are also used to assess the quality of a realized sample and to decide on any corrections necessary. The distribution (of a combination) of background variables in a realized sample is compared with the population characteristics from official data. Deviations from the known population distributions can be corrected by appropriate weights." Demographic and socio-economic background variables describe social and cultural concepts of societies and social structures. Besides the three classical variables – sex, age, and education – the number of demographic and socio-economic variables needed to determine relationships between attitudes and social characteristics depends on the research question.[3] The three classical background variables, together with occupation, labour force status and employment status, ethnicity and migration background, family- and household structure, and "total net household income", serve to typify the respondents and to describe the social context in which they act.

In the social sciences, income data are used as independent socio-economic variables to explain individual differences in social position within a society. Because disposable income is an indicator of the purchasing power that enables people to satisfy their social needs, net household income, and the disposable per capita income of each household member derived from it, determine standards of living and lifestyles in stratified societies (see Lepsius, 1974; Lepsius, 1993: 156ff.). Individuals' different positions in the social network of relationships constitute social inequality. In the words of Hradil (2006: 195f.), "social inequality" refers to the living conditions (working conditions, income, material wealth, education attainment, etc.) that allow some people to achieve generally shared goals of a "good life" (for example health, security, wealth, respect) better than others.

A social science concept of income as a background variable must succeed in explaining differences in behaviour, attitudes, orientations, and membership of forms of social organisation observed across groups of respondents. For this reason, an instrument for the measurement of income for social scientific analyses must capture the "more versus less" of net household income and the "top versus bottom" in the

[3] Depending on the research interest, it is useful to select further background variables, for example religion, political orientation, marital status, dwelling, health, nutrition, body mass index, etc.

income distribution, and must allow income differences to be examined. Cross-national comparisons have revealed that not only income levels and income differences are of relevance but also the types of income and the different sources from which income is drawn. This is due to the fact that income classes also differ in terms of the composition of the income they receive, and that countries differ in terms of "income packaging" (see Rainwater, Rein and Schwartz, 1987). Hence, in 1968 Gösta Carlsson (p. 189) counted the source and level of the income among the definitive variables of social stratification. It later emerged that income composition also affects the quality of the information provided by respondents.

# 3 Measurement instrument "total net household income" used in the European Social Surveys 2002, 2004, and 2006

The European Social Survey is an academically driven cross-national social survey that has been conducted every two years across Europe since 2002. By now, the survey collects data about attitudes, beliefs and behaviour of populations in over thirty nations. The aims of the ESS are to study stability and change in social structure, opinions and attitudes of citizens in Europe, and to provide indicators of citizens' perceptions and judgments of aspects of their societies and social and political life. To achieve comparability in the operationalisation of the ESS across countries, the Core Scientific Team produces a detailed project specification. These common standards and requirements apply to sample selection, questionnaire translation, and to all methods and processes during data collection and processing, and documentation.

The household income measure is part of the question module aimed at creating a socio-demographic profile of each respondent (Section F). This module includes questions on household composition, sex, age, type of residential area, education and occupation of the respondent and of his or her partner and parents, trade union membership, and marital status. The main parts and questions in this section are stable over the consecutive survey rounds. However, the household income question was changed in Round 4 of the ESS, which was fielded in 2008. The corresponding instructions to the national coordinators who supervise the fieldwork in their respective countries were also modified.

## 3.1 Measurement instrument used in the European Social Survey 2002

The questionnaire used in Round 1 of the ESS 2002 featured two questions designed to measure household income. The first question (F29) asked the respondent to state the main source of income in his or her household; the second question (F30) aimed

to identify the income category to which the household's "total net income" belonged. To this end, the respondent was requested to "add up the income from all sources". However, in this pan-European survey, the randomly selected respondents were not given any detailed background information or explanations of the questions. Hence, it was not clear to them which income – and whose income – they should add up. No explanation or definition of "net income" was provided. Respondents were not given any help in recalling the various possible types and sources of income accruing to the household either. Because the interviewees are randomly selected from among all the members of the household aged 16 or over, and only the target person is interviewed, respondents' knowledge of the financial situation of the household as a whole varies depending on the cohort to which they belong and to their position in the household or their relationship to the main income earner/recipient. In 2002, the ESS question about the main source of income in the household read:

> "F29 CARD 55 Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household? Please use this card" (European Social Survey, 2002a: 49).

The showcard listed seven types of income:

> "Wages or salaries;
> Income from self-employment or farming;
> Pensions;
> Unemployment / redundancy benefit;
> Any other social benefits or grants;
> Income from investment, savings, insurance or property;
> Income from other sources" (European Social Survey, 2002b: CARD 55).

The respondent was then asked to calculate the total net income of the household and to assign it to one of the income categories presented on the showcard.

> "F30 CARD 56 Using this card, if you add up the income from all sources, which letter describes your household's total net income? If you don't know the exact figure, please give an estimate. Use the part of the card that you know best: weekly, monthly or annual income" (European Social Survey, 2002a: 47).

The ESS Project Instructions for 2002 featured the following interviewer instruction regarding the definition of "net income". However, this information was not intended for the respondent.

"At HINCTNT you should obtain the total net income of the household from all sources, that is, after tax. Income includes not only earnings but state benefits, occupational and other pensions, unearned income such as interest from savings, rent, etc. We want figures after deductions of income tax, national insurance, contributory pension payments and so on. The questions refer to cur-

| CARD 56 | | | | |
|---|---|---|---|---|
| YOUR HOUSEHOLD INCOME | | | | |
| | **Approximate WEEKLY** | **Approximate MONTHLY** | **Approximate ANNUAL** | |
| J | Less than €40 | Less than €150 | Less than €1800 | J |
| R | €40 to under €70 | €150 to under €300 | €1800 to under €3600 | R |
| C | €70 to under €120 | €300 to under €500 | €3600 to under €6000 | C |
| M | €120 to under €230 | €500 to under €1000 | €6000 to under €12000 | M |
| F | €230 to under €350 | €1000 to under €1500 | €12000 to under €18000 | F |
| S | €350 to under €460 | €1500 to under €2000 | €18000 to under €24000 | S |
| K | €460 to under €580 | €2000 to under €2500 | €24000 to under €30000 | K |
| P | €580 to under €690 | €2500 to under €3000 | €30000 to under €36000 | P |
| D | €690 to under €1150 | €3000 to under €5000 | €36000 to under €60000 | D |
| H | €1150 to under €1730 | €5000 to under €7500 | €60000 to under €90000 | H |
| U | €1730 to under €2310 | €7500 to under €10000 | €90000 to under €120000 | U |
| N | €2310 or more | €10000 or more | €120000 or more | N |

(European Social Survey, 2002b: CARD 56)

> rent level of income or earnings or, if that is convenient, to the nearest tax or other period for which the respondent is able to answer. The respondent is given a showcard that enables them to choose between their weekly, monthly or annual income, whichever they find easiest. They will then give you the letter that corresponds to the appropriate amount. This system is designed to reassure the respondent about the confidentiality of the information they are giving." (European Social Survey, 2002c: 21).

In the first three rounds of the ESS 2002, 2004, and 2006, the central coordinators of the survey prescribed a common and uniform system of income categories for all participating countries for use in the response to the income question.

Table 1 illustrates the survey outcomes of ESS 2002 for selected countries. The countries shown here represent different income distributions. German respondents use mainly the answer categories 4 to 9 (6,000 to 60,000 euros). The majority of respondents in the United Kingdom declare the "total net household income" using the income ranges from category 4 to 10 (6,000 to 90,000 euros) with a high frequency on the ninth income category (36,000 - 60,000 euros). The survey participants in Italy answer with the income categories 3 to 9 (3,600 to 60,000 euros). In Luxembourg the upper income categories 5 to 11 (12,000 to 120,000 euros) are used. Respondents in Portugal answer the income question with the lower categories 2 to 6 (1,800 to 24,000 euros). Interviewees in Finland use the categories 2 to 7, 11 and 12 (1,800 to 30,000, 90,000 – 120,000 and 120,000 and more euros).

**Table 1:** Distribution of "total net household income" in ESS 2002 for selected countries

| | Germany | | United Kingdom | | Italy | |
|---|---|---|---|---|---|---|
| | % | valid % | % | valid % | % | valid % |
| 1. Up to 1,800 | .47 | .59 | .49 | .58 | .44 | .81 |
| 2. 1,800-3,600 | .77 | .98 | 1.20 | 1.42 | 1.09 | 2.02 |
| 3. 3,600-6,000 | 1.40 | 1.77 | 2.63 | 3.12 | 3.31 | 6.15 |
| 4. 6,000-12,000 | 6.51 | 8.23 | 10.60 | 12.54 | 10.10 | 18.73 |
| 5. 12,000-18,000 | 12.58 | 15.90 | 9.44 | 11.17 | 10.13 | 18.79 |
| 6. 18,000-24,000 | 13.37 | 16.90 | 9.35 | 11.06 | 9.24 | 17.13 |
| 7. 24,000-30,000 | 13.25 | 16.75 | 7.75 | 9.17 | 7.95 | 14.75 |
| 8. 30,000-36,000 | 9.83 | 12.42 | 8.19 | 9.69 | 3.56 | 6.60 |
| 9. 36,000-60,000 | 14.15 | 17.89 | 19.12 | 22.62 | 5.77 | 10.70 |
| 10. 60,000-90,000 | 4.95 | 6.25 | 9.42 | 11.14 | 1.50 | 2.78 |
| 11. 90,000-120,000 | 1.12 | 1.42 | 3.42 | 4.05 | .43 | .81 |
| 12. 120,000 and more | .72 | .91 | 2.91 | 3.45 | .39 | .73 |
| 77. Refusal | 13.59 | | 5.59 | | 33.89 | |
| 88. Don't know | 7.31 | | 9.88 | | 12.20 | |
| 99. No answer | | | | | | |
| Total % | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| valid N | | 2308.92 | | 1734.50 | | 650.73 |
| missing N | | 610.13 | | 317.53 | | 556.27 |

| | Luxembourg | | Portugal | | Finland | |
|---|---|---|---|---|---|---|
| | % | valid % | % | valid % | % | valid % |
| 1. Up to 1,800 | .04 | .07 | 1.48 | 2.26 | 4.15 | 4.63 |
| 2. 1,800-3,600 | .62 | .97 | 5.61 | 8.58 | 12.50 | 13.96 |
| 3. 3,600-6,000 | .41 | .64 | 11.13 | 17.02 | 14.20 | 15.86 |
| 4. 6,000-12,000 | .68 | 1.07 | 16.74 | 25.59 | 13.60 | 15.19 |
| 5. 12,000-18,000 | 3.06 | 4.82 | 11.79 | 18.03 | 14.35 | 16.02 |
| 6. 18,000-24,000 | 7.66 | 12.07 | 7.88 | 12.04 | 11.30 | 12.62 |
| 7. 24,000-30,000 | 10.41 | 16.40 | 3.93 | 6.01 | 15.35 | 17.14 |
| 8. 30,000-36,000 | 8.74 | 13.77 | 2.60 | 3.98 | 3.00 | 3.35 |
| 9. 36,000-60,000 | 17.44 | 27.48 | 2.81 | 4.30 | .75 | .84 |
| 10. 60,000-90,000 | 10.13 | 15.96 | .88 | 1.34 | .35 | .39 |
| 11. 90,000-120,000 | 3.26 | 5.14 | .32 | .49 | 1.70 | 13.96 |
| 12. 120,000 and more | 1.02 | 1.61 | .23 | .35 | 8.70 | 15.86 |
| 77. Refusal | 18.38 | | 8.18 | | .05 | |
| 88. Don't know | 14.72 | | 15.03 | | 12.50 | |
| 99. No answer | 3.42 | | 11.40 | | 14.20 | |
| Total % | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| valid N | | 984.95 | | 988.22 | | 1791.0 |
| missing N | | 566.70 | | 522.78 | | 209.0 |

Source: ESS 2002, own calculations

## 3.2 Comparison of the results for "total net household income" from the European Social Survey 2002 and the 2001 European Community Household Panel

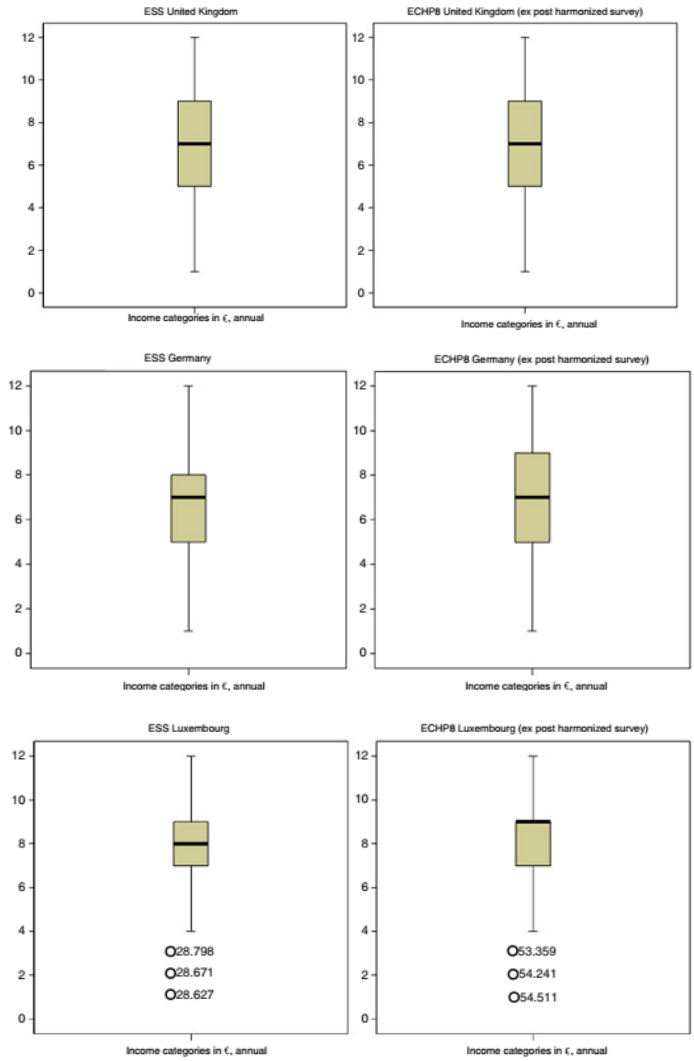
The European Community Household Panel (ECHP) collects all the types of household income that may occur in the country in question; all household members aged 15 years or over are interviewed. All respondents are asked in detail about their income during the year prior to the interview. Hence, in the course of their involvement in the panel, respondents become experts on their personal monetary situation. The field instrument, which is designed as a personal questionnaire, lists all possible sources of monetary income. Each member of the household is able to recall and state all individually applicable income types from various sources during the interview. The 34 types of income listed by the ECHP take up over 16 pages in the personal questionnaire. In addition to the individual questionnaire for each member of the household aged 15 and older, a household questionnaire is administered to a reference person in the household who is assumed to be able to provide reliable information about income that cannot be assigned to individual members but rather accrues to the household as a whole. The household questionnaire covers 19 types of income, for example, "social assistance payment, non-cash assistance from the welfare office, income from renting property, inheritance of property or capital, a gift or lottery winnings" (European Commission/Eurostat, 2000: 25-27). Because this survey of the income situation of the household and its members is so comprehensive and detailed, the data from Wave 8 of the ECHP (2001), with the income reference year 2000, can be used as a reference for the measurement of "total net household income" in ESS 2002.

We have recoded the ECHP income values into the income categories used in the ESS. The images on the left of Figure 1 are graphical representations of the distribution of responses across income categories in ESS 2002 for the respective countries. The images on the right of Figure 1 show the grouped income distribution in ECHP8 2001.

In the case of the United Kingdom, both data sources yield the same income distribution. A slight deviation is apparent in the case of Germany, while marked differences between the two statistics are evident in the case of Luxembourg.

The national income distributions from the 8th wave of the ECHP – divided into groups, each of which contains 5% of the population – constitute the second step in the comparison of the "total net household income" data of the two surveys (see Table 2). They are sorted into the income categories used as response options by the ESS. This step highlights the need to adapt the response categories of the income question to the concrete national income situation.

**Figure 1:** Distribution of "total net household income" using ESS 2002 categories: Comparison of ESS and ECHP for selected countries




Source: Hoffmeyer-Zlotnik and Warner, 2014: 147

In Germany, the 15th to the 19th 5-percent percentile of the ECHP are to be found in the 9th ESS 2002 income category (36,000 – 60,000 euros); the 10th ECHP8-2001 5-percent percentile, whose upper threshold corresponds to the median of the income distribution, is in the 7th ESS 2002 income category (24,000 – 30,000 euros).

According to ECHP8-2001, only the wealthiest 5% of Portuguese households have a "total net household income" of over 36,000 euros. In Luxembourg, the ninth ESS1-2002 income category (36,000 – 60,000 euros) covers the ECHP's income distribution from the ninth to the 15th 5-percent percentile. The bottom 5% of the population in the ECHP income distribution for Luxembourg have a net household

income of between 12,000 and 18,000 euros (the 5th ESS1-2002 category), whereas in Portugal the median (the 10th 5-percent percentile) is to be found in the fourth income category (6,000 – 12,000 euros).

**Table 2:** Distribution of the ECHP8 5-percent percentiles across the 12 ESS 2002 income categories for selected countries

| ESS Income category | Germany | United Kingdom | Italy | Luxembourg | Portugal | Finland |
|---|---|---|---|---|---|---|
| | No. of the ECHP8 5% Percentile | | | | | |
| 1. Up to 1,800 | --- | --- | --- | --- | --- | --- |
| 2. 1,800-3,600 | --- | --- | --- | --- | 1-2 | --- |
| 3. 3,600-6,000 | --- | --- | 1 | --- | 3-5 | --- |
| 4. 6,000-12,000 | 1-2 | 1-2 | 2-5 | --- | 6-11 | 1-3 |
| 5. 12,000-18,000 | 3-5 | 3-5 | 6-10 | 1 | 12-15 | 4-7 |
| 6. 18,000-24,000 | 6-8 | 6-7 | 11-13 | 2-3 | 16-17 | 8-10 |
| 7. 24,000-30,000 | 9-12 | 8-10 | 14-16 | 4-6 | 18 | 11-12 |
| 8. 30,000-36,000 | 13-14 | 11-12 | 17 | 7-8 | 19 | 13-15 |
| 9. 36,000-60,000 | 15-19 | 13-17 | 18-19 | 9-15 | --- | 16-19 |
| 10. 60,000-90,000 | --- | 18-19 | --- | 16-18 | --- | --- |
| 11. 90,000-120,000 | --- | --- | --- | 19 | --- | --- |
| 12. 120,000 and more | --- | --- | --- | --- | --- | --- |

Source: Hoffmeyer-Zlotnik and Warner, 2014: 148

Overall, the household income of the respondents in Germany and Luxembourg is distributed across six and seven income categories respectively. However, depending on the average national income, the distribution across income categories varies significantly across countries. In Portugal, the top four response categories of the ESS should not have been used in the survey; in Luxembourg the lower four categories should have been left blank.

# 4 The improved income measure: "Total net household income" in the European Social Surveys 2008, 2010, and 2012

The modifications to the income questions in the fourth round of the ESS 2008 affected the framing of the questions, the response categories, and the showcards:

> "F31 CARD 72 Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household? Please use this card." (European Social Survey, 2008b: 59)

The modified showcard features separate response options for "income from self-employment (excluding farming)" and "income from farming":

> "Wages or salaries;
> Income from self-employment (excluding farming)
> Income from farming
> Pensions
> Unemployment/redundancy benefit
> Any other social benefits or grants
> Income from investment, savings, insurance or property
> Income from other sources" (European Social Survey, 2008b: 59)

The text of the "total household income" question gives the respondent an indication of what is meant by "net", as it specifies "after tax and compulsory deductions".

> "Using this card, please tell me which letter describes your household's total income, after tax and compulsory deductions, from all sources? If you don't know the exact figure, please give an estimate. Use the part of the card that you know best: weekly, monthly or annual income." (European Social Survey, 2008b: 60)

Since the fourth round of the ESS, each participating country has drafted its own showcard. The response categories are based on the deciles of the actual household income distribution in the country in question.

In a note on the drafting of the decile income showcard, the ESS coordinators provide the following instructions to those responsible for running the survey in each country: "An income showcard should be devised with approximate **weekly, monthly and annual amounts**. You should use **ten income range categories, each corresponding broadly to DECILES OF THE ACTUAL HOUSEHOLD INCOME RANGE in your country**. These figures should be derived from the best available source for your country. The data source used should match the requirement of the question i.e. deciles of household income for all households (not for example average households or just households with children). Using the median income as

the reference point, 10 deciles should be calculated with the median itself at the top of the fifth decile (Category F). The figures should not appear to be too exact. Minor rounding can be employed to achieve this if necessary" (European Social Survey, 2008b: 60; see also European Social Survey, 2008c: 17).

| CARD 73 | | | | |
|---|---|---|---|---|
| YOUR HOUSEHOLD INCOME | | | | |
| | **Approximate WEEKLY** | **Approximate MONTHLY** | **Approximate ANNUAL** | |
| J | Weekly equivalent | Monthly equivalent | Income corresponding to that held by 10% of households with lowest income (0-10%) | J |
| R | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (11-20%) | R |
| C | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (21-30%) | C |
| M | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (31-40%) | M |
| F | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (41-50%) | F |
| S | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (51-60%) | S |
| K | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (61-70%) | K |
| P | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (71-80%) | P |
| D | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (81-90%) | D |
| H | Weekly equivalent | Monthly equivalent | Income corresponding to that held by next 10% of households (91-100%) | H |

Source: European Social Survey, 2008b: CARD 73

Twenty-six countries participated in Round 4 of the ESS 2008. Figure 2.1 shows the country-specific distributions of the responses across the 10 income categories. The medians of the income distributions of 15 countries lay in the fifth or sixth income category, thereby fulfilling the ESS requirement quoted above. In six countries, the medians were in a category above the sixth income category, while in five countries the medians of the distribution were in a category below the fifth category. In most of the countries, 50% of the observed household incomes were spread over four or five categories. As the response options are based on the deciles of the national income distribution, we expected that approximately 50% of the surveyed population would use five categories to answer the ESS income question. However, in the Czech Republic half of the respondents used only two income ranges, while in Portugal three categories were used by the surveyed persons.

Figure 2.2 shows the country-specific distributions of the responses in Round 5 of the ESS, which was fielded in 2010. As in the previous round, 26 countries participated. The medians of the income distributions in 22 countries lay in the fifth or sixth

income category. In one country the median was in a category above the sixth income category, while in three countries the medians of the distribution were in a category below the fifth category.

**Figure 2.1:** Country-specific distributions of responses across the ten income categories in ESS 2008

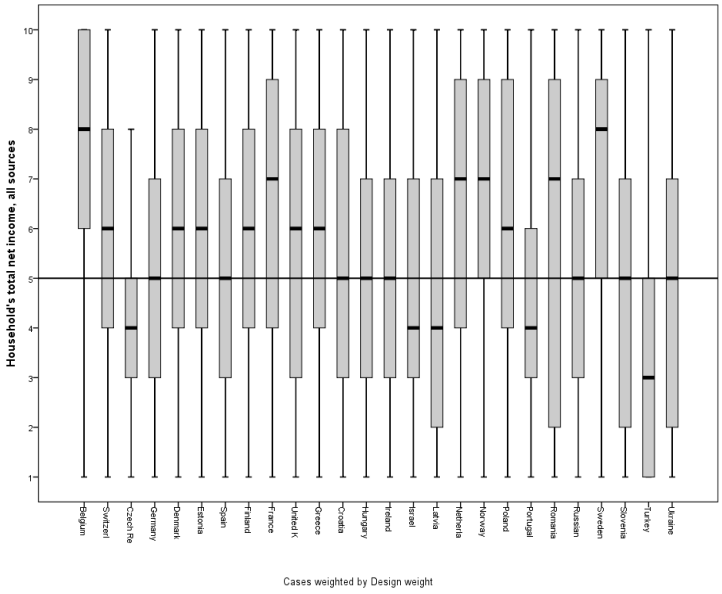



**Figure 2.2:** Country-specific distributions of responses across the ten income categories in ESS 2010

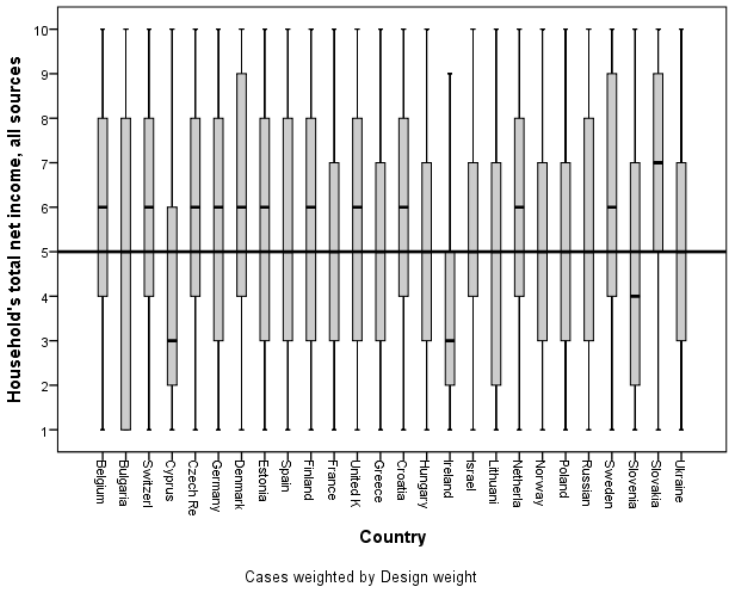
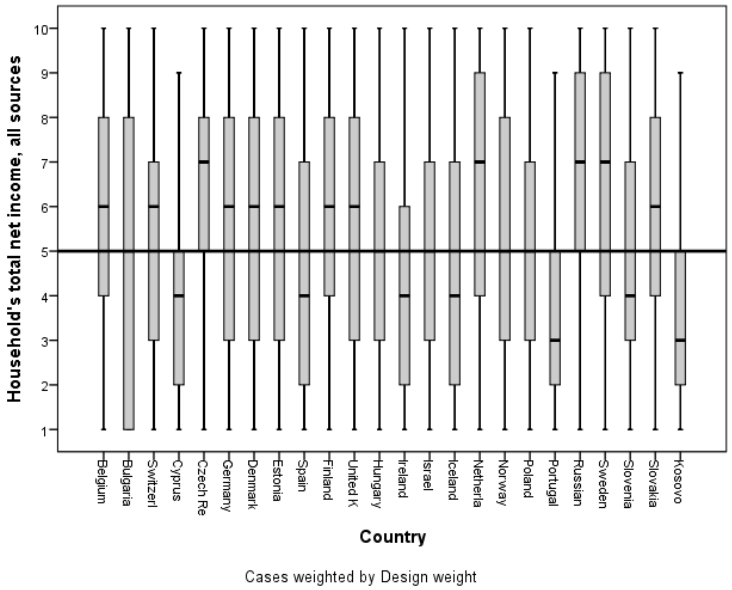


**Figure 2.3:** Country-specific distributions of responses across the ten income categories in ESS 2012




Source for figures 2.1–2.3: ESS 2008, 2010, 2012, own calculations. The solid horizontal line shows the expected median

Figure 2.3 illustrates the distributions in 2012. Data from 24 countries were available for the analysis. The medians of 13 countries lay in the fifth or sixth income category. In four countries, the medians were in a category above the sixth income category, while in seven countries the medians of the distribution were in a category that was lower than expected. In Bulgaria, seven income categories were used by 50% of the surveyed population, whereas in Portugal, Kosovo, and the Czech Republic only three options were used.

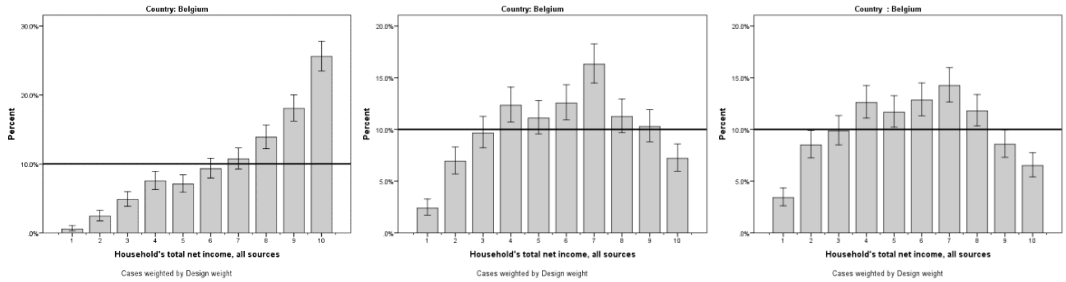
As the income categories on the showcard for the income question correspond to the deciles of the actual household income range, it is to be expected that in a representative survey with a probabilistic sample each response category will be selected by approximately 10% of the survey population.

As can be seen from the countries presented by way of example in Figure 3, our expectation was fulfilled in some cases, but not in others. In ESS 2008, for example, each income category was chosen by almost 10% of respondents in Denmark, Estonia, Finland, France, the United Kingdom, Croatia, Poland, and Slovenia. However, medium-sized deviations from the expected decile distribution were observed in the case of Switzerland, Germany, Spain, Greece, Hungary, the Netherlands, Norway, the Ukraine, and Ireland, where the middle income categories were more strongly represented than expected. In 2008, large deviations from the decile distribution were observed in Belgium, the Czech Republic, Latvia, Portugal, Romania, Russia, Sweden, and Turkey.

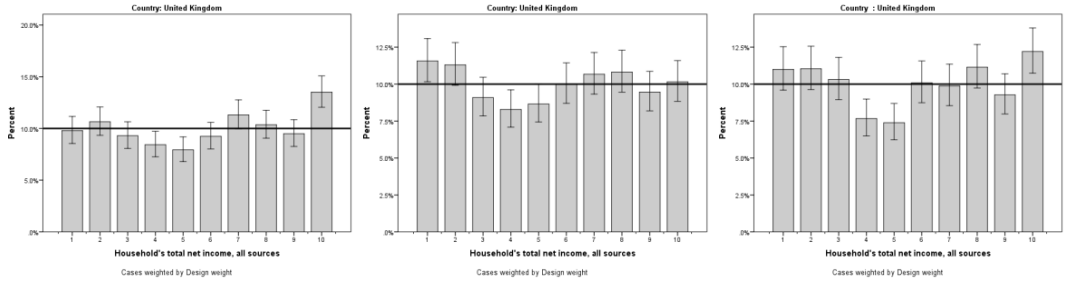
**Figure 3:** Distributions across the income categories in ESS 2008, ESS 2010 and ESS 2012 for selected countries

ESS 2008 ESS 2010 ESS 2012

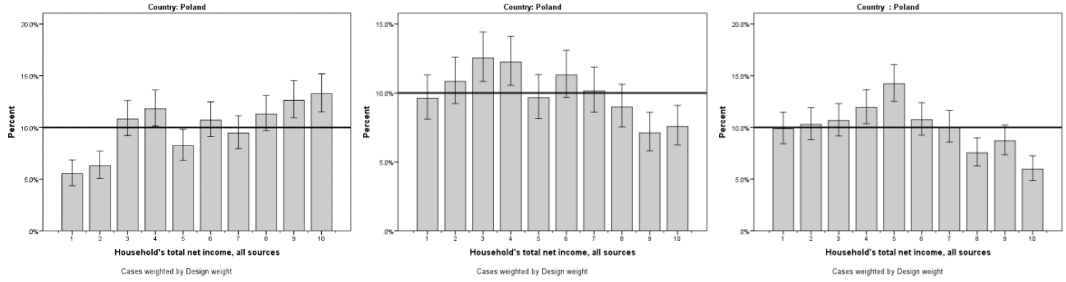
1. Belgium



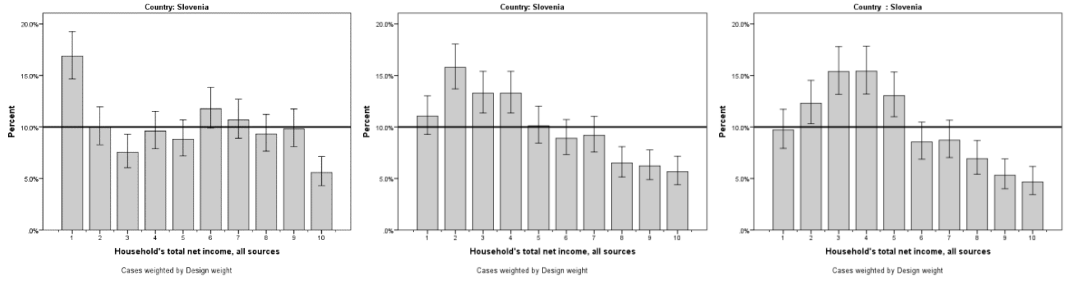

2. United Kingdom




3. Poland

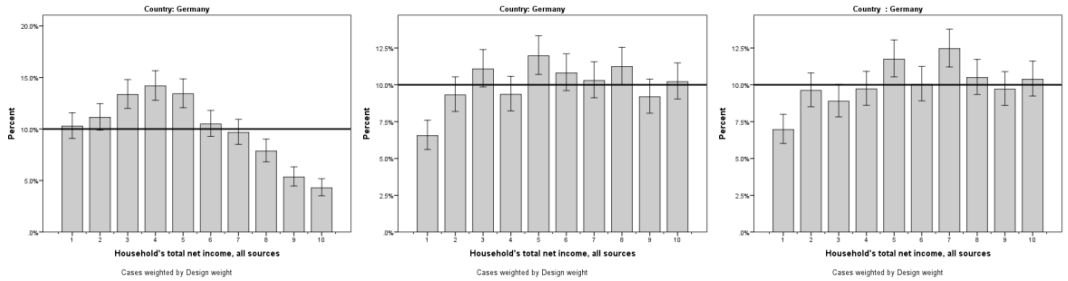



4. Slovenia



ESS 2008 ESS 2010 ESS 2012

5. Germany




Source: ESS 2008, 2010, 2012, own calculations. The solid horizontal line shows the expected 10% responses, error bars = 95% confidence interval

In the Belgian ESS 2008, the two highest deciles show large deviations from the expected 10% mark (Figure 3, row 1). The highest income category starts at 35,000 euros. However, 33,731 euros is the upper threshold of the 60% decile of EU-SILC[4] in Belgium (Table 2). Therefore, considerably more than the expected 10% of the respondents in Belgium chose the ninth and tenth income categories during the ESS interview. The lower income categories were not used to the expected extent by the ESS 2008 respondents. The EU-SILC reports the threshold of the lowest decile at 12,012 euros, which corresponds to the fourth income category on the showcard used by Belgium in ESS 2008 (Tables 3 and 4).

**Table 3:** EU-SILC 2008 "total disposable household income" decile thresholds in Euros for Belgium

| lowest 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | highest 90% |
|---|---|---|---|---|---|---|---|---|
| 12,012 | 15,191 | 18,741 | 22,837 | 27,683 | 33,731 | 40,012 | 47,386 | 59,951 |

Source: EU-SILC USER DATABASE Version of 01-08-11, own calculations

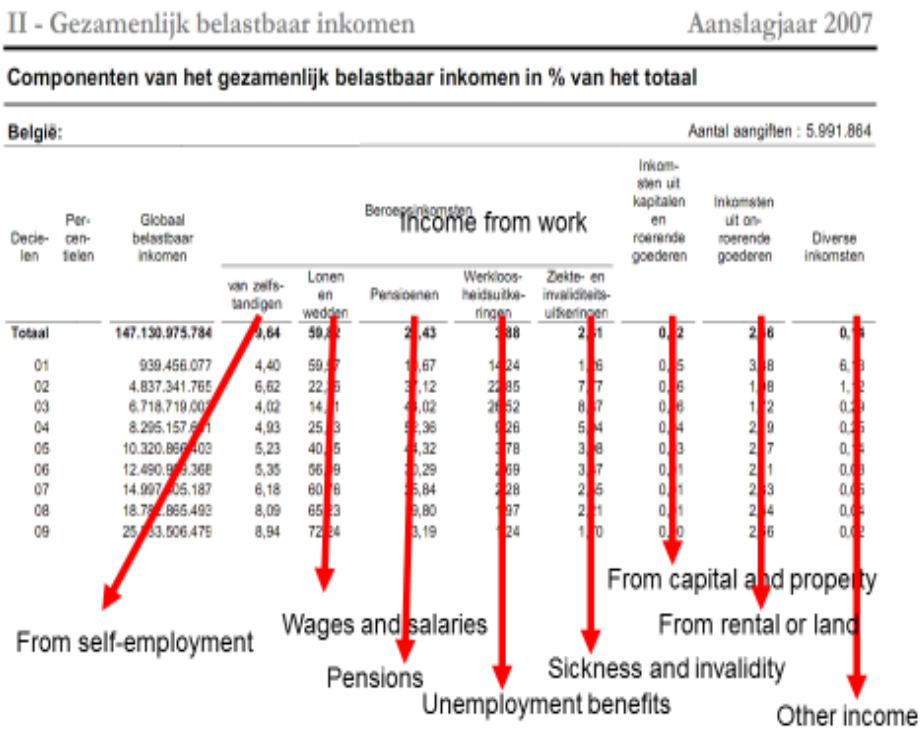
[4] The European Union Statistics on Income and Living Conditions (EU-SILC) are comparable multidimensional micro-data on income, social inclusion, and living conditions. They cover objective and subjective aspects of these themes in both monetary and non-monetary terms for both households and individuals. They are used to monitor the progress of the Europe 2020 strategy – in particular its headline target on poverty reduction, which provides information on income, poverty, social exclusion, housing, labour, education and health. http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/introduction#

**Table 4:** Income distribution in Belgium according to the tax register and the income brackets used in the Belgian ESS 2008

| Deciles | Total taxable net income from register | Average tax paid in % | (Total taxable net income from register)-(Average tax paid) | Rounded net income as appeared on Showcard 72 |
|---|---|---|---|---|
| 1 | 4,909 | 0 | 4,909.000 | Less then 5,000 € |
| 2 | 9,677 | 1.5 | 9,531,845 | 5,000 € to 10,000 € |
| 3 | 12,001 | 2.3 | 11,724,977 | 10,000 € to 12,000 € |
| 4 | 14,860 | 7.9 | 13,686,060 | 12,000 € to 14,000 € |
| 5 | 18,139 | 12.5 | 15,871,625 | 14,000 € to 16,000 € |
| 6 | 21,816 | 17.9 | 17,910,936 | 16,000 € to 18,000 € |
| 7 | 26.457 | 21.2 | 20,848,116 | 18,000 € to 21,000 € |
| 8 | 34,146 | 24,3 | 25,848,522 | 21,000 € to 26,000 € |
| 9 | 47,834 | 27.5 | 34,679,650 | 26,000 € to 35,000 € |
| 10 | >47,834 | >27.5 | >34,679,650 | 35,000 € or more |

Source: European Social Survey 2008d: 3

**Table 5:** Components of taxable income in Belgium




Source: Algemene Directie Statistiek en Economische Informatie, 2009: 36, translation by the authors

The ESS4-2008 Survey Documentation (European Social Survey, 2008d: 38) reported that the income response categories for Belgium were calculated on the basis of "total taxable net income" data from the "Tax statistics for revenues of 2004".

The responses in Belgium gave rise to major deviations from the expected 10% mark in all ten response categories. In Belgium, taxable income is made up of wages and salaries, income from self-employment, pensions, unemployment benefit, sickness and disability benefit, income from the rental of property and land, income from investments, and income from property and other sources (Table 5). However, because the ESS measures "total net household income", and many components of household income – for example public and private transfers – are not subject to tax, it is obvious that the lower response categories in Belgium were either not used at all or hardly used.

The Polish showcard for the answers to the income question in ESS4-2008 is based on the income distribution of the Polish Household Budget Surveys. In ESS4-2008, the lower two income categories were underrepresented and did not reach the 10% mark; the highest two income deciles were used more often than expected (Figure 3, row 3). We observed no significant changes in the answers to the income questions in the subsequent rounds.

The ESS 2008 respondents fulfilled expectations in the United Kingdom, as only the highest decile was overrepresented and less than 10% chose the middle income categories (Figure 3, row 2). Response behaviour was similar in the subsequent ESS rounds. The showcards containing the income categories are based on the data from the Family Resources Surveys.

Figure 3, row 4 shows that in ESS 2008 in Slovenia the 10% requirement was fulfilled in the case of the majority of the ten income categories. The lowest category was overrepresented and the highest answer option was not used by the expected 0% of the eligible respondents (Table 6).

**Table 6:** Slovenia: Observed and expected responses across the ten income categories

| | 2008 | | | 2010 | | | 2012 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Observed N** | **Expected N** | **Residual** | **Obs. N** | **Exp. N** | **Resid.** | **Obs. N** | **Exp. N** | **Resid.** |
| **1** | 172 | 102.0 | 70.0 | 119 | 107.7 | 11.3 | 96 | 96.7 | -.7 |
| **2** | 102 | 102.0 | .0 | 170 | 107.7 | 62.3 | 121 | 96.7 | 24.3 |
| **3** | 77 | 102.0 | -25.0 | 143 | 107.7 | 35.3 | 152 | 96.7 | 55.3 |
| **4** | 98 | 102.0 | -4.0 | 143 | 107.7 | 35.3 | 144 | 96.7 | 47.3 |
| **5** | 90 | 102.0 | -12.0 | 109 | 107.7 | 1.3 | 130 | 96.7 | 33.3 |
| **6** | 120 | 102.0 | 18.0 | 96 | 107.7 | -11.7 | 85 | 96.7 | -11.7 |
| **7** | 109 | 102.0 | 7.0 | 99 | 107.7 | -8.7 | 85 | 96.7 | -11.7 |
| **8** | 95 | 102.0 | -7.0 | 70 | 107.7 | -37.7 | 64 | 96.7 | -32.7 |
| **9** | 100 | 102.0 | -2.0 | 67 | 107.7 | -40.7 | 49 | 96.7 | -47.7 |
| **10** | 57 | 102.0 | -45.0 | 61 | 107.7 | -46.7 | 41 | 96.7 | -55.7 |
| **Total** | 1020 | | | 1077 | | | 967 | | |

Source: ESS, 2008, 2010, 2012, own calculations

However, in the case of all other categories, the 10% benchmark was within the 95% confidence interval of equal distribution over the 10 income ranges. This changed in 2010 and 2012. In 2008, Slovenia showed small deviations from the 10% equal distribution over the income categories. The income categories were based on the 2007 Census figures provided by the Statistical Office of the Republic of Slovenia (European Social Survey, 2008d: 253). In 2010 and 2012, the deviations were deemed to be medium, and the reference statistics for the creation of income categories were the EU-SILC (European Social Survey, 2010: 27; 2012: 27).

In ESS 2008 in Germany, categories 3, 4, 5, and 7 to 10, and their corresponding 95% confidence intervals, did not respect the 10% benchmarks (see Figure 3, row 5). For the 2010 and 2012 rounds of the ESS, the German ESS coordinators changed the statistical basis for the creation of the income categories. They switched from the 2003 Einkommens- und Verbrauchsstichprobe (Income and Consumer Survey)[5] (European Social Survey 2008 e: 30 f.) as reference statistics to the 2008 and 2011 Mikrozensus (a 1% population census)[6] respectively (European Social Survey, 2010b: 13 and European Social Survey, 2012b: 12), with the result that the deviations from the equal distribution over ten categories in 2010 and 2012 were deemed to be small (see Table 7).

**Table 7:** Germany: Observed and expected responses across the ten income categories

| | 2008 | | | 2010 | | | 2012 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Observed N** | **Expected N** | **Residual** | **Obs. N** | **Exp. N** | **Resid.** | **Obs. N** | **Exp. N** | **Resid.** |
| **1** | 269 | 228.6 | 40.4 | 170 | 226.9 | -56.9 | 208 | 255.3 | -47.3 |
| **2** | 289 | 228.6 | 60.4 | 229 | 226.9 | 2.1 | 271 | 255.3 | 15.7 |
| **3** | 322 | 228.6 | 93.4 | 259 | 226.9 | 32.1 | 246 | 255.3 | -9.3 |
| **4** | 329 | 228.6 | 100.4 | 224 | 226.9 | -2.9 | 265 | 255.3 | 9.7 |
| **5** | 297 | 228.6 | 68.4 | 273 | 226.9 | 46.1 | 306 | 255.3 | 50.7 |
| **6** | 226 | 228.6 | -2.6 | 241 | 226.9 | 14.1 | 252 | 255.3 | -3.3 |
| **7** | 201 | 228.6 | -27.6 | 231 | 226.9 | 4.1 | 302 | 255.3 | 46.7 |
| **8** | 159 | 228.6 | -69.6 | 246 | 226.9 | 19.1 | 246 | 255.3 | -9.3 |
| **9** | 108 | 228.6 | -120.6 | 189 | 226.9 | -37.9 | 224 | 255.3 | -31.3 |
| **10** | 86 | 228.6 | -142.6 | 207 | 226.9 | -19.9 | 233 | 255.3 | -22.3 |
| **Total** | 2286 | | | 2269 | | | 2553 | | |

Source: ESS, 2008, 2010, 2012, own calculations

In 2008, four countries used the EU-SILC as the basis for calculating the household income deciles; fourteen countries calculated the household income ranges on the basis of other survey data; and eight countries derived the income deciles from population registers or census data (Table 8).

[5] The German Income and Consumer Survey is a quota sample realized every five years.

[6] The German 1% population census is a random sample of households. Participation in the survey is mandatory for the household members selected.

**Table 8:** Sources of the income distribution used to create the ESS income categories on the showcards, and deviations from the 10% frequencies per category

| | 2008 | | 2010 | | 2012 | |
|---|---|---|---|---|---|---|
| **country** | **deviation** | **source** | **deviation** | **source** | **deviation** | **source** |
| Belgium | large | register | medium | SILC | small | SILC |
| Bulgaria | | | large | register | large | register |
| Croatia | small | survey | medium | survey | | |
| Switzerland | medium | survey | small | survey | small | survey |
| Cyprus | | | large | SILC | large | SILC |
| Czech Rep | large | SILC | medium | SILC | medium | SILC |
| Germany | medium | survey | small | census | small | census |
| Denmark | small | register | medium | register | small | register |
| Estonia | small | SILC | medium | SILC | small | survey |
| Spain | medium | survey | medium | survey | large | survey |
| Finland | small | survey | small | survey | small | survey |
| France | small | census | small | survey | | |
| UK | small | survey | small | survey | small | survey |
| Greece | medium | SILC | medium | SILC | | |
| Hungary | medium | survey | medium | survey | small | survey |
| Israel | medium | survey | large | survey | small | survey |
| Kosovo | | | | | large | survey |
| Latvia | large | SILC | | | | |
| Lithuania | | | large | survey | | |
| Netherlands | medium | register | small | register | medium | register |
| Norway | medium | register | small | register | small | register |
| Poland | small | survey | small | survey | medium | survey |
| Portugal | large | survey | | | large | SILC |
| Romania | large | survey | | | | |
| Russian Fed | large | survey | small | survey | large | survey |
| Sweden | large | register | large | survey | medium | survey |
| Slovakia | | | medium | SILC | small | SILC |
| Slovenia | small | census | medium | SILC | medium | SILC |
| Turkey | large | survey | | | | |
| Ukraine | medium | survey | small | survey | | |
| Ireland | medium | survey | large | SILC | large | SILC |
| Iceland | | | | | large | survey |

Note: A deviation is deemed to be **large** if at least one response category deviates by at least 10 percentage points from the expected 10 percent mark. A deviation is considered to be **medium** if at least one response category deviates by at least 5 percentage points from the ten percent mark. Deviations of 2.5 percentage points from the expected 10 % share are deemed to be **small**.

Source: Data Documentation Reports of ESS, 2008, 2010 2012 available as Survey Documentation Report from http://www.europeansocialsurvey.org/data/round-index.html [accessed 21 March 2014], own calculations

# 5. Conclusion

In this article we have focused on the ESS question about the "total net household income". In Hoffmeyer-Zlotnik and Warner, 2014 (228 ff.), we presented a proposed survey instrument of our own for obtaining information about this background variable.

As average income levels and income distributions differ in various types of European countries, the response categories must be adapted to the national income situation of the surveyed country. The quality of the responses to the survey question about "total net household income", and ultimately the quality of the survey data obtained, depends on the quality of the reference statistics from which the household income ranges for the answers are derived.[7]

The example from Belgium shows the increase in the quality of income data collected in the ESS that can be achieved by changing the reference statistics. In 2008, the national coordinators used the tax register with limited income information to design the income categories on the showcard, which resulted in large deviations from the expectation that each category would be chosen by 10% of respondents. In 2010 and 2012, Belgium's national coordinators used the income data provided by EU-SILC, thereby reducing the deviations from the expected responses.

The German example shows a different case. In 2008, the national coordinators based the construction of the "total net household income" categories on the Income and Consumer Survey (Einkommens- und Verbrauchsstichprobe). In the 2010 and 2012 rounds of the ESS they used the Mikrozensus (a 1% population census) as reference statistics for the income ranges offered to the respondents. As a result, a larger number of response categories were closer to the expected 10% benchmark.

This leads us to conclude that the quality of income measurements for comparative social surveys depends on the elaborateness with which the response options are created. This, in turn, depends on the quality of the reference statistics about the "total net household income" used to determine the income ranges on the showcard presented to the respondents. Other possible effects are related to the use of scales in surveys: country specific practices in survey design and the culture specific answer behaviour of survey participants. The stability of the national income distribution over time may have an impact on the quality of income measurement in social surveys.

[7] Other survey errors also have an impact on the objectivity, reliability, validity, and measurement quality of the total net household income data. They include non- or undercoverage of the sampling frame, systematic unit-non-response, non-random item-non-response, national or cultural attitudes towards highly sensitive survey topics, opinions about the data protection and the privacy of personal information, the interpersonal relation between interviewer and interviewee, translation error, inadequate weighting and wrong extrapolation factors, faulty replacements and weak imputations of missing data, ultimately doubtful and dubious equivalences of the information across cultures and countries, etc. (cf. Groves and Lyberg, 2010 for survey errors in general, for the problem of equivalences in cross cultural and cross national surveys cf. Johnson, 1998, and in particular for errors in comparative survey research cf. Braun, 2003).

To date, a total of seventy-six national datasets have been collected over all rounds of the ESS (see table 9). In nineteen cases, the income categories were based on the national EU-SILC data. Only three (15.7%) of these datasets show small deviations from the national SILC data. Forty datasets used national surveys; 17 (42.5%) of these datasets show small deviations. Thirteen countries used register data and four countries used census data to determine the response options for the ESS interviews.

**Table 9:** Reference statistics used to create the income categories, by deviation

| | | Deviation | | | Total |
|---|---|---|---|---|---|
| | | Large | Medium | Small | |
| National Source | SILC (n) | 7 | 9 | 3 | 19 |
| | (%) | 36.8 | 47.3 | 15.7 | 100.0 |
| | Survey (n) | 11 | 12 | 17 | 40 |
| | (%) | 27.5 | 30.0 | 42.5 | 100.0 |
| | Register (n) | 4 | 4 | 5 | 13 |
| | (%) | 30.7 | 30.7 | 38.4 | 100.0 |
| | Census (n) | 0 | 0 | 4 | 4 |
| | (%) | 0.0 | 0.0 | 100.0 | 100.0 |
| Total (n) | | 22 | 25 | 29 | 76 |
| (%) | | 28.9 | 32.8 | 38.1 | 100.0 |

Note: A deviation is deemed to be **large** if at least one response category deviates by at least 10 percentage points from the expected 10 percent mark. A deviation is considered to be **medium** if at least one response category deviates by at least 5 percentage points from the ten percent mark. Deviations of 2.5 percentage points from the expected 10 % share are deemed to be **small**.
Source: ESS Data Documentation Reports and own calculations

# 6. Recommendations

Data from national sources have to fulfil quality criteria if they are to be used as reference statistics to establish the income categories for the survey question on household income (see Ehling and Körner 2007: 9 f.).

The first requirement is the relevance of the source data for the household's income information – that is, the degree to which the reference data meet the needs of the survey. This is expressed in the following instruction to the national coordinators of the ESS: "You should use ten income range categories, each corresponding broadly to DECILES OF THE ACTUAL HOUSEHOLD INCOME RANGE in your country. These figures should be derived from the best available source for your country. The data source used should match the requirement of the question …". (European Social Survey, 2008a: 60; see also European Social Survey,

2008b: 17). The formulated question adds "… all household members and any income" and income "from all sources".

The second quality component for reference data is accuracy. Accuracy describes the correspondence of the income values in the reference statistics with the (unknown) true income distribution in the entire population.

The next quality criterion is timeliness, which refers to the time between the date of the reference data on income and the date of the survey measurement of income in the field. The requirement mentioned by the ESS is "the actual household income" at the time of the interview. However, it is evident that there is a time lag between the date of the official income information and the start of the survey fieldwork.

The fourth quality domain is comparability. The concepts of "total net household income" applied in the reference statistics and in the survey instruments must be as similar as possible. Deviations between the reference data and the outcomes of the survey are comparable if they are not due to the tools applied during data collection. A comparison of the income distribution yielded by the survey with the income distribution of the reference statistics is meaningful.

The information from the reference statistics must be coherent with the intended survey measurements. The main issue with regard to coherence is that both the reference statistics and the survey use the most similar approaches, classifications, and methodological standards possible when operationalising "total net household income".

Finally, the accessibility and clarity of the reference statistics are important. The researchers who prepare, design, and organise the fieldwork of the surveys must be able to find the income information necessary to create the income categories for the survey question about "total net household income". This presupposes not only access to the individual data from the reference statistics, clear documentation about the data production, and metadata about the source of the reference statistics. The major elements of clarity are the corresponding quality reports by the providers of the reference statistics.

We recommend that,

first, the central coordinators of the ESS define and communicate minimum threshold values for each of the aforementioned quality domains.

Second, the national coordinators have the best knowledge about the statistical sources available in their country. A close collaboration with experts on the national income situation of households is indispensable to select the reference statistics and to draft the showcard for the question about "total net household income". Together they decide about the best available reference statistics taking the guidelines from the central coordinators into account. Ideal are national surveys collecting income data with similar technology and instruments like ESS, e.g. as a module in the annual Labour Force Surveys. Today, only a small number of countries include "total net household income" in the national Labour Force Surveys.

Third, the national coordinators responsible for the design of the questionnaire and the measurement instrument implemented in the survey should report their decision and justify the selection of the reference statistics. The documented national deviations from the guidelines of the central coordinators and the reference statistics on the income become transparent to the users of ESS.

The users of ESS data have now an insight into the quality of the income measurement during the interviews. If these recommendations are followed, we expect more and better cross-country comparisons using the socio-demographic background variable "total net household income".

## Acknowledgements

The authors wish to thank CEPS/INSTEAD, Esch, Alzette for financial support.

## References

[1] Algemene Directie Statistiek en Economische Informatie (2009): *Levensstandaard Fiscale statistiek van de inkomens. Aanslagjaar 2007 – Inkomens van 2006.* Brussel economie.fgov.be/nl/binaries/Inkomens%20brochure%20B%20aanslagjaar%202007_tcm325-78903.pdf [accessed 21 March 2014]

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

[3] Braun, M. (2003): Errors in Comparative Survey Research: An Overview. In: J. A. Harkness, van de Vijver, J.R. F. and Mohler, P. Ph. (Eds.) *Cross-cultural survey method*: 137–142. Hoboken, NJ: Wiley.

[4] Braun, M. and Mohler, P. Ph. (2003): Background Variables. In: J. A. Harkness, van de Vijver, J.R. F. and Mohler, P. Ph. (Eds.): *Cross-cultural survey method*: 101–116. Hoboken, NJ: Wiley.

[5] Carlsson, G. (1968): Ökonomische Ungleichheit und Lebenschancen. In: Glass, David V.; König, R. (Eds) Soziale Schichtung und soziale Mobilität. *KZfSS Sonderheft 5*.

[6] Ehling, M. and Körner, T. (Eds.) (2007): *Handbook on Data Quality Assessment Methods and Tools*. Wiesbaden: Statistisches Bundesamt.

[7] European Social Survey (2002a): *Source Questionnaire. Round 1, 2002*. ESS Document Date 01-08-02. http://www.europeansocialsurvey.org/docs/round1/fieldwork/source/ESS1_source_main_questionnaire.pdf [accessed 21 March 2014].

[8] European Social Survey (2002b): *Source Showcards. Round 1, 2002*. http://www.europeansocialsurvey.org/docs/round1/fieldwork/source/ESS1_source_showcards.pdf [accessed 21 March 2014].

[9] European Social Survey (2002c): *Project Instructions (PAPI). Round 1, 2002*. ESS Document date 15/07/02. http://www.europeansocialsurvey.org/docs/round1/fieldwork/source/ESS1_source_project_instructions.pdf [accessed 21 March 2014].
[10] European Commission/Eurostat (2000): *ECHP 2001 Wave 8. Variable List DOC. PAN 159/00*.
[11] European Social Survey (2008a): *The European Social Survey, Source Questionnaire Amendment 03 (Round 4, 2008/9)*. ESS Document Date 08-08-08. http://ess.nsd.uib.no/ess/round4/fieldwork.html. [accessed 18 October 2012].
[12] European Social Survey (2008b): *ESS 2008 Data Protocol Edition 1.2, October 2008*. ESS4_data_protocol_e1.2.pdf. http://ess.nsd.uib.no/ess/round4/. [accessed 18 October 2012].
[13] European Social Survey (2008c): *Data Documentation Report. Ed. 5.1. Appendix A5* http://www.europeansocialsurvey.org/docs/round4/survey/ESS4_appendix_a5_e05_0.pdf [accessed 18 December 2013]
[14] European Social Survey (2008d): *ESS4 - 2008 Documentation Report Edition 5.2* http://www.europeansocialsurvey.org/data/download.html?r=4 [accessed 18 October 2014]
[15] European Social Survey (2010a): *Data Documentation Report. Ed. 3.0. Appendix A2* http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_appendix_a2_e03_0.pdf [accessed 18 December 2013]
[16] European Social Survey (2010b): *ESS5 - 2010 Documentation Report Edition 3.1 Appendix A2, Income, ESS5 - 2010 ed. 3.0* http://www.europeansocialsurvey.org/data/download.html?r=5 [accessed 18 October 2014]
[17] European Social Survey (2012a): *Data Documentation Report. Ed. 1.3. Appendix A2* http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_appendix_a2_e01_1.pdf [accessed 18 December 2013]
[18] European Social Survey (2012b): *ESS6 – 2012 Documentation Report Edition 2.0 Appendix A2, Income, ESS5 - 2012 ed. 2.0* http://www.europeansocialsurvey.org/data/download.html?r=6 [accessed 18 October 2014]
[19] Groves, R. and Lyberg, L. (2010): Total survey error: Past, present and future. *Public Opinion Quarterly*, **74 (5)**, 849-879.
[20] Hoffmeyer-Zlotnik, J. H.P. and Warner, U. (2006): Methodological Discussion of the Income Measure in the European Social Survey Round 1; in: *Metodoloski zvezki* Vol. **3**, No. 2: 289-334.
[21] Hoffmeyer-Zlotnik, J. H.P. and Warner, U. (2014): *Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research*. Dordrecht, Heidelberg, New York, London: Springer Science + Business Media.
[22] Hradil, S. (2006): *Die Sozialstruktur Deutschlands im internationalen Vergleich*. 2. Auflage. Wiesbaden: VS-Verlag.

[23] Johnson, T. P. (1998): Approches to Equivalence in Cross Cultural and Cross-National Survey Research. In: Harkness, Janet (Ed.): *Cross-Cultural Survey Equivalence*. ZUMA-Nachrichten Spezial 3. Mannheim: ZUMA

[24] Lepsius, R. M. (1993): *Demokratie in Deutschland. Soziologisch-historische Konstellationsanalysen. Ausgewählte Aufsätze*. Göttingen: Vandenhoeck & Ruprecht.

[25] Lepsius, R. M. (1974): Sozialstruktur und soziale Schichtung in der Bundesrepublik Deutschland. In: Löwenthal, R. and Schwarz, H.-P. (eds). *Die zweite Republik. 25 Jahre Bundesrepublik Deutschland – eine Bilanz*. Stuttgart: Seewald.

[26] Lynn, P. (2001): *Developing Quality Standards for Cross-National Survey Research: Five Approaches*. ISER Working Papers Number 2001-21 www.iser.essex.ac.uk/…er_working_papers/2001-21.pdf [accessed 1 October 2014]

[27] ONS (2013): *Guidelines for Measuring Statistical Output Quality, Version 4.1*. ONS, London. http://www.ons.gov.uons method / guide- /method-quality/quality/guidelines-for k/- -quality/index.html [accessed 1 October 2014]

[28] Rainwater, L., Rein, M. and Schwartz, J. (1987): *Income Packaging in the Welfare State: A Comparative Study of Family Income*. Oxford University Press.

[29] Smith, T. W. and Yang-Chih F. (2014): *The Globalization of Surveys.* GSS Cross-national Report No. 34. publicdata.norc.org/gss/documents/CNRT/CNR34.pdf [accessed 1 October 2014]

[30] Statistics Canada (1998): *Statistics Canada Quality Guidelines*. 3rd edition, Ottawa: Statistics Canada.

[31] United Nations Economic Commission for Europe (2011): *Canberra Group Handbook on Household Income Statistics*. Second Edition. Geneva: United Nations

[32] Warner, U., (2012): 'Total Net Household Income' as Demographic Standard Variables for Social Survey; in: Hoffmeyer-Zlotnik, J. H.P. and Warner, U. (Eds.): *Demographic Standards for Surveys and Polls in Germany and Poland – National and European Dimension*. GESIS-Schriftenreihe Band 10, 187-208. Köln: GESIS – Leibniz-Institut für Sozialwissenschaften.

Metodološki zvezki, Vol. 12, No. 2, 2015, 111-122

# Environmental, Generation and Policy Determinants of Feed-in Tariff: a Binary Pooling and Panel Analysis

Antonio A. Romano, Giuseppe Scandurra, Alfonso Carfora [1]

**Abstract**

In this paper we analyze the key-factors behind the adoption of the Feed-in Tariff. We propose and test two regression models for binary data: a pooling specification and a panel one. We employ a comprehensive sample of 60 countries with distinct economic structures over the period 1980–2008. Economics, environmental and generation factors are used as regressors and results demonstrate that these factors are relevant for the policy decision to adopt the Feed-in Tariff. Furthermore, the panel specification appears a better specification, in a such heterogeneous context, than the classical pooled specification.

## 1 Introduction

Nowadays, about 80% of electric power is generated by fossil sources (coal, gas and oil), but there are growing global concerns regarding the lack of sustainability of these forms of electricity generation that bring into question their use in long–term energy development strategies. During the last decade in order to respond to energy–related challenges such as climate change, air pollution, volatility in fossil fuel prices, and a growing demand for electricity, countries have multiplied recourse to renewable energy sources (RES). RES are becoming increasingly important in the energy generation mix of countries, because they can reduce global climate change, the dependence on imported fossil fuels, but they also promote the local economic development. For this reasons, governments have adopted a wide variety of grants and /or incentives aimed to support renewable energy investments.

There is a wide range of policies being used to support renewable energy development around the world, including Feed–in Tariff (FiT), renewable portfolio standards (RPS), economic tools, distributed generation measures and disclosure and green marketing measures. FiT and RPS are two of the most popular policy instruments. A FiT program typically guarantees that customers who own a FiT-eligible renewable electricity generation facility, such as a roof-top solar photo-voltaic system, will receive a set price from their utility for all of the electricity they generate and provide to the grid.

A brief overview of the main policy instruments being used to promote RES can be found

[1] Department of Management Studies and Quantitative Methods, University of Naples "Parthenope" - via Generale Parisi, 13 - Naples (Italy)- I - 80132; alfonso.carfora@uniparthenope.it

in a recent report by the *United Nations Environment Programme* (UNEP, 2012). The report produced by the UNEP highlights that FiT is the most applied national policy instrument to promote RES.

There is plenty of literature about RES and policy instruments and, for sake of simplicity, it can be divided in two main topics: the former, in which authors describe and assess FiT policies and the latter in which policies are included in the key factors of investments in RES. Lesser and Su (2008) propose an innovative two–part FiT, consisting of both a capacity payment and a market-based energy payment, which can be used to meet the renewable policy goals of regulators. They find that the proposed two–part tariff design draws on the strengths of traditional FiT, relies on market mechanisms, is easy to implement, and avoids the problems caused by distorting wholesale energy markets through above-market energy payments. Del Rio (2012) builds a theoretical framework for dynamic efficiency analysis and assess the dynamic efficiency properties of the different design elements of FiTs. He shows that several design elements can have a significant impact on the different dimensions of dynamic efficiency. Particularly relevant design elements in this context are technology-specific fixed–tariffs, poor prices, reductions of support over time for existing plants, long duration of support and support falling on consumers. Dong (2012) analyzes the effectiveness of FiT and renewable portfolio standard (RPS) in the development of wind generation. He finds that FiT policies have a positive effect on RES development while RPS policies have a negative effect. Other authors (Islam and Meade, 2013) using data from Ontario, where a generous FiT is available to households generating electricity from solar panels, measure household level preferences for panels and use these preferences along with household characteristics to predict adoption time intentions. Hsu (2012) employs a system dynamics model in order to develop a simulation for assessing which policy, or combination of policies, promoting solar PV applications has the greatest economic benefit. The simulation period is from 2011 to 2030. He finds that FiT price or subsidy is a good approach. Stokes (2012) presents a case study of Ontario's FiT policies between 1997 and 2012 to analyze how the political process affects renewable energy policy design and implementation.

Other studies examine the drivers of RES development and include the policy grants in order to quantify the effect of these instruments in the promotion of RES. Among these the most interesting (Menz and Vachon, 2006; Carley, 2009) study the renewable investments in the USA, the former with a regression into countries and the latter using a panel regression while Marques et al. (2010) analyze the drivers promoting renewable energy in European countries and find that lobbies of traditional energy source and $CO_2$ emission restrain renewable deployment. Evidently, the need for economic growth suggests an investment that supports, but does not replace, the before installed capacity. Romano and Scandurra (2011) analyze the investments in RES in low carbon and high carbon economies using a panel dataset. More recently the same Authors study the key factors promoting the investments in RES in a panel dataset of Petroleum Exporting Countries (OPEC) members (Romano and Scandurra, 2014 and Romano et al., in press), and the role of economic growth as driver of the FiT adoption (Romano et al., 2015). In the first case, lack of grants and/or incentives to promote the installations of new renewable power plants has been considered a limit for the sustainable development of the OPEC countries, in the second it came to light as the economic growth is one of the main driver of the FiT adoption.

The aim of the paper is to identify the determinants driving a country's choice of adopting FiT policy. We address this issue using a static panel probt model estimated over a pooling and panel specification as longitudinal analysis can improve the results. For this reason we use a comprehensive dataset of 60 countries with distinct economic and social structures as well as different levels of economic development in the years between 1980 and 2008.The sample includes OECD, South American, Asian and African countries. This dataset can be helpful to assess the effect that macroeconomic variables have in order to suggest to policymakers the need to adopt the FiT. The organization of the paper is as follows: Section 2 describes data and scope of the work while Section 3 analyzes the models proposed and reports the empirical results discussing about the policy implications. Section 4 contains a comparative analysis between the pooling and panel specification. Concluding remarks are given in Section 5.

## 2 Data and scope of the work

The empirical analysis is based on a large dataset of 60 countries with different economic and social structures. We use annual data from 1980 to 2008 obtained from the *U.S. Energy Information Administration* (EIA) and *International Renewable Energy Agency* (IRENA). All the countries in the sample have a share of electricity generated by RES but around 40% of our sample does not adopt the FiT. In this way we include in the sample also countries that generate electricity with RES but do not adopt this policy instruments to promote new RES power plants.
The variables used limit the major economic, generation, and environmental factors from which investment decisions are originated and influence the policymakers. As in the UNEP report (UNEP, 2012) we classify the explanatory variables in four homogeneous factors:

- Environmental (total $CO_2$ emission from energy consumption);
- Economics (per capita electricity consumption; GDP per capita; energy security);
- Generation (share of non–hydroelectric renewable generation; share of nuclear generation; share of fossil generation);
- Policy (adoption of Kyoto protocol).

Among environmental factors we include the total carbon dioxide emissions from the consumption of energy that capture the environmental degradation due to economic development. The International Energy Agency evaluates that $CO_2$ from energy represents about three quarters of the anthropogenic GHG emissions for Annex I[2] countries, and

[2]The Annex I Parties to the 1992 UN Framework Convention on Climate Change (UNFCCC) are: Australia, Austria, Belarus, Belgium, Bulgaria, Canada, Croatia, the Czech Republic, Denmark, Estonia, European Economic Community, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Monaco, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom and United States. See www.unfccc.int.

over 60% of global emissions. This percentage presents high variability by country, due to different national structures (IEA, 2013). The choice to include in the regressors the total $CO_2$ ($lnCO_2$) instead of the per capita $CO_2$ seems appropriate because international agreements binding targets for reducing total greenhouse gas (GHG) emissions. In fact, $CO_2$ emissions are one of the main factors of the greenhouse effect and actions have been embraced to reduce it. Obviously, the carbon dioxide emission is a proxy for environmental degradation. The expected result is a significant positive effect. The more the emissions are, the more should be the probability to adopt the FiT.

Economic factors includes GDP per capita (lnGDP), per capita consumption of energy (lnCons) and the energy security of supply (Import). The GDP is one of the most important economic indicators. It is commonly assumed that richer countries are able to better promote investments in RES, employing various forms of grants and incentives. GDP is also related to energy consumption, which is considered a proxy for economic development of the country (Toklu et al., 2010). The increasing energy consumption leads policymakers to build new power plants based on renewable sources, better technology or economic structural changes. In the literature this is associated to a higher strength in environmental degradation. A similar argument can be applied to energy security, approximated by the degree of dependence on foreign supplies of electricity. As known, the power grid interconnections allow a constant exchange of electricity between countries. The need to increase their share of generation (and to reduce electricity dependence) could increase the probability of adopting policies to support the RES development.

Among generation factors we include the share of non–hydroelectric renewable generation (ShRENNH), i.e. the ratio between non–hydro renewable generation and total net electricity generation that can be also considered a proxy of investments in RES (Romano and Scandurra, 2014; in press). The effect is expected positive. Countries are encouraged to increase the share of electricity generation from RES. We include the non–hydroelectric generation because, generally, FiT is mainly related to promote the investments in these source (mainly photo-voltaic and wind generation) rather than hydroelectricity.

Much of the world's electricity is generated thermally using non–renewable (fossil) fuels. Thermal generation (ShTHER) has both a high environmental impact and presents increasing generation costs. Despite the growth of non–fossil energy, the share of fossil fuels within the world energy supply is relatively unchanged over the past 40 years. In 2011, fossil sources accounted for 82% of the global TPES. Generation of electricity and heat worldwide relies heavily on coal, the most carbon–intensive fossil fuel. Countries such as Australia, China, India, Poland and South Africa produce over two-thirds of their electricity and heat through the combustion of coal (IEA, 2013). Finally, among generation factors we also include the share of nuclear generation (ShNUC). The nuclear energy generation occurs only in some countries, mainly rich countries and it is $CO_2$ free.

As policy factors we include a dummy variable that indicates the adoption of Kyoto protocol. The Kyoto Protocol is an international agreement linked to the United Nations Framework Convention on Climate Change, which commits industrialized countries (as a group) to curb domestic emissions by about 5% relative to 1990 by the 2008–12 first commitment period. Alongside the agreement to negotiate a new climate agreement by 2015, 38 countries have agreed to take commitments under a second commitment period of the Kyoto Protocol to begin in 2013.

Clearly, not all aspects of a complex phenomenon like the decisions to adopt some grants

for investments in renewable energy can be disclosed in the present work. Some critical issues, such as the reprogramming of the energy plan, problems related to the population, the environmental impacts of new power plants, are not taken into account but they are factors that can affect the decisions.

# 3 The models

## 3.1 The static probit model

Let us define $Y_t \in \{0, 1\}$ the binary time series at time $t$. The linear predictor of the static probit model (Greene, 2003) is defined as:

$$\pi_t = \alpha + x'_{t-k}\beta. \tag{3.1}$$

with t = 1,...,T

It assumes that the expected value of $Y_t$ conditionally on information at time *t-k* is given by

$$E(Y_t) = p_t = \Phi(\pi_t)$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution.

The main feature of the static model is that it does not consider lagged dependent variables as regressors.
For this reason only exogenous lagged variables are used as regressors. Probit models are often presented in a dynamic specification (Estrella and Mishkin, 1998; Kauppi and Saikkonen, 2008; Nyberg, 2010; De Luca and Carfora, 2014). Here, the choice of the static specification is due to the specific nature of the outcome variable, not compatible with a dynamic specification, considering that a country's choice to adopt the tariff represents a medium term strategy in environmental policy. Following eq. (3.1) the static probit model is:

$$\begin{aligned}\pi_t = \alpha + \beta_1 lnCO_{2t-1} + \beta_2 lnCons_{t-1} + \beta_3 lnGDP_{t-1} + \beta_4 Imports_{t-1} + \\ + \beta_5 ShRENNH_{t-1} + \beta_6 Kyoto_t + \beta_7 ShTHER_{t-1} + \beta_8 ShNUC_{t-1}.\end{aligned} \tag{3.2}$$

We calculate maximum likelihood estimators of the parameters and listed them together with their standard errors in Table 1.

The coefficients are in line with the expected results. Among the significant coefficients, except for the electricity consumption, that (as expected) is negatively linked to the outcome variable, the increase in one of the explanatory variables is directly related to the increase in the probability to adopt the FiT. This is an important result that suggests that countries consider the FiT as an useful instrument to promote the RES to reduce the carbon emissions (especially after the subscription to the Kyoto protocol). Moreover there is a direct relationship between the FiT adoption and tendency of the GDP. The coefficient of the net imports is not significant. This is due to the presence, of countries that depend on others for their electricity consumption. This is an interesting issue in the analysis of the phenomenon and for this reason we do not drop the variable by the model-specification. Energy security is a relevant aspect in the electricity models but it does not relevant in

**Table 1:** Estimates, standard errors, p-values of the Static Probit Model.

| Variables | Estimates | Std Errors | P-values |
|---|---|---|---|
| $Constant$ | -9.274 | 1.295 | 0.000 |
| $lnCO_{2t-1}$ | 0.344 | 0.036 | 0.000 |
| $lnCons_{t-1}$ | -0.196 | 0.114129 | 0.087 |
| $lnGDP_{t-1}$ | 0.686 | 0.145 | 0.000 |
| $Imports_{t-1}$ | 0.043 | 0.027 | 0.110 |
| $ShRENNH_{t-1}$ | 3.862 | 1.001 | 0.000 |
| $ShTHER_{t-1}$ | -0.094 | 0.184 | 0.611 |
| $ShNUC_{t-1}$ | -0.168 | 0.320 | 0.600 |
| $Kyoto_t$ | 1.122 | 0.096 | 0.000 |

the decision to adopt FiT. Furthermore, by the analysis of the coefficients we observe that a relationship between the outcome variable and thermal and nuclear generation factors, being this type of policy unconnected with these factors, does not exist. Maybe electricity demand, is mainly supplied by the traditional sources of energy which are still independent from the innovative policies. Also these issues appear interesting and lead us to let these variables in the model specification.

## 3.2 The panel probit model

We improve the traditional probit model attempting to explain the probability that a country will adopt the FiT in terms of exogenous variables and individual characteristics too. This approach includes determinants related to the heterogeneity in the dataset following a panel specification of the traditional static model.

The panel model is the following:

$$\pi_{i,t} = \omega + x'_{i,t-k}\beta + u_i + z_{i,t}. \tag{3.3}$$

where $z_{i,t}$ represents the error term and $u_i$ denotes a country-specific random effect and $u_i + z_{i,t} = \epsilon_{i,t}$ . The assumption on $u_i$ is that it is i.i.d drawn from a univariate normal distribution or $u_i \sim N(0; \sigma_u)$ (Vella and Verbeek, 1998).

As in the pooling specification (3.1), it occurs that:

$$E(Y_{i,t}) = p_t = \Phi(\pi_{i,t})$$

The individual specific unobserved effects are uncorrelated with the independent variables and with the error terms. For the estimation of the coefficients of the panel binomial models (logit and probit) are often used fixed effects estimator (Jakubson, 1991) or procedures based on instrumental variables (see, e.g., Hausman and Taylor, 1981; Amemiya and McCurdy, 1986). Pinheiro and Bates (1995) review several methods to calculate the parameters in a generalized linear mixed effect model maximum likeklhood (ML) procedures. In this work the maximum likelihood method proposed by Vella and Verbeek(1998) has been used in order to estimate the parameters in (3.4) (derived directly by the 3.2),

and the unobserved heterogeneity $\sigma_u$ of the random effects probit model. Due to the presence of time-invaring variables (like the energy imports or share of nuclear generation, constantly equal to zero for several countries), the choice of a random model appears as the most appropriate specification. In fact, using a fixed-effects model that eliminates the effects through time-demeaning, time-constant variables will be drop out too. Therefore, the use of time-demeaning variables is equivalent to introducing a full set of individual dummies, that, taken together, would be collinear with any time-invariant ones.

$$\begin{aligned}\pi_{i,t} = \omega + \beta_1 lnCO_{2i,t-1} + \beta_2 lnCons_{i,t-1} + \beta_3 lnGDP_{i,t-1} + \beta_4 Imports_{i,t-1} + \\ + \beta_5 ShRENNH_{i,t-1} + \beta_6 Kyoto_{i,t} + \beta_7 ShTHER_{i,t-1} + \beta_8 ShNUC_{i,t-1}\end{aligned} \tag{3.4}$$

In Table 2 we summarize the results obtained by the maximum likelihood estimation of the parameters under the panel specification (3.4).

**Table 2:** Estimates, standard errors, p-values of the estimated Panel Probit Model.

| Variables | Estimates | Std Errors | P-values |
|---|---|---|---|
| $Constant$ | -9.656 | 1.392323 | 0.000 |
| $lnCO_{2t-1}$ | 0.349 | 0.03926 | 0.000 |
| $lnCons_{t-1}$ | -0.236 | 0.123001 | 0.055 |
| $lnGDP_{t-1}$ | 0.741 | 0.15586 | 0.000 |
| $Imports_{t-1}$ | 0.057 | 0.034049 | 0.095 |
| $ShRENNH_{t-1}$ | 3.229 | 1.075743 | 0.003 |
| $ShTHER_{t-1}$ | -0.168 | 0.194903 | 0.388 |
| $ShNUC_{t-1}$ | -0.160 | 0.334758 | 0.633 |
| $Kyoto_t$ | 0.423 | 0.181158 | 0.020 |
| $\sigma_u$ | 1.114 | 0.28393 | 0.000 |

The coefficients of the panel model are in line, both in signs and in the intensities, with those estimated in the model under the pooling specification. Under the panel specification, also the coefficient related to net imports becomes significant event though at 10 per cent level. This is due to the use of the panel random effect (RE) model specification and hence of the correct error covariance that, eliminating the endgeneity between error terms and regressors, returned a more consisted estimator. Moreover, the significance of the standard deviation ($\sigma_u$) of the effects confirms the presence of the heterogeneity between the countries and validates the choice of the random model.

# 4 Comparison of model specifications

The selected variables used to describe the key-factors underlying the decision of a country to adopt the Feed-in Tariff seem to be adequate to describe the phenomenon and the results, both of the models under the pooling and under the panel specification. Moreover, results appear to be consistent with the theoretical implications recognized in literature.

The aim of this work is also to assess if, in presence of an heterogeneous set of countries observed in different years, a panel specification can be a valid tool to obtain more appropriate estimates. Moreover we try to improve the results capturing, once isolated, the variability of the country specific effects that is part of information that in a pooling specification should be turned up in the residual component. To reach this goal we have calculated, both for the pooling and for the panel model a set of most commonly used fitting measures. These are:

1. The value of $maximized\ loglikelihood$ of the model
2. The $Akaike\ Information\ Criterion$ - AIC (Akaike,1974)

$$AIC(r) = -2\ell + 2r$$

where $\ell$ is the maximized log-likelihood of the model and $r$ denotes the number of parameters estimated.

3. The $pseudo - R^2$ proposed by Mc Fadden(1973)

$$R^2_M = 1 - \frac{\ell_1}{\ell_0}$$

where $\ell_1$ is the maximized log-likelihood of the considered model and $\ell_0$ is the maximum of the log-likelihood function under a Null model without any explanatory variable.

And, finally, as index of predictive accuracy,

4. The $Mean\ Absolute\ Error$ (MAE):

$$MAE = \left( \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} |Y_{i,t} - \hat{\pi}_{i,t}|}{NT} \right), \tag{4.1}$$

where $\hat{\pi}_{i,t}$ is the fitted probabilities under the panel model to adopt the FiT. With $t$ starting in 1981 and ending in 2008 (T) for a total of 28 years and $i = 1, ..., 60$ corresponds to the 60 countries under study.

**Table 3:** Measures of comparison of models.

| Specification | Maximum Loglikelihood | AIC | $R^2_M$ | MAE |
|---|---|---|---|---|
| $Pooling$ | -500.77 | 1019.54 | 0.33 | 0.18 |
| $Panel$ | -487.45 | 994.90 | 0.35 | 0.18 |
| LRT test | $\chi^2$ statistic: 26.64 (0.000) | | | |

Table 3 reports the different measures used to compare the goodness of fit, the predictive accuracy of different estimated models and results of Likelihood ratio tests (LRT) to test the pooling restriction of the static probit model. While using MAE, there is no difference between the two models, other results indicate the panel as the better specification. Moreover, results of the Likelihood Ratio Test (LRT), lead us to reject the null hypothesis of indifference between the two models in favour to the panel parameterization ($\chi^2$ statistic: 26.64; p-value: 0.000). Thus, a panel model including also individual components is the better model to evaluate the determinants of a country's choice to adopt the FiT.

# 5 Conclusions

This study has two essential aims: *i)* identify the variables related to the decisions of a country to adopt an incentive as a Feed-in Tariff to promote the investments in RES and *ii)* find an appropriate model specification to extrapolate informative contents from an heterogeneous set of countries.
Regarding the former, both the models return positive and similar responses, reaching the main goal of the analysis. Almost all selected variables are significant on the decision to appeal to the incentive. In particular, the absolute value of the $CO_2$ emissions assumes importance both as impact on the population of the level of pollution and as emissivity measure of the efficiency of the generation process in a country. In fact, the collapse in the price of RES observed in recent years, put forward the latter as a viable alternative to expensive plant system upgrades necessary to cut $CO_2$ emissions. The increasing probabilities to promote the electricity generation from RES that emerge from this model, strengthen the evidence of the widespread use of energy policies oriented towards generating conversion from RES. In the same way should the results related to Kyoto variable be interpreted. Moreover, the economic growth is certainly one of the major cause of the FiT adoption. As GDP level increases and living conditions improve, countries try to introduce some policy incentives for the RES. The results related to the electricity consumption in both models are also in line with the theoretical outcomes. From our results, it emerges a plausible support of the causality between reduction in the electricity consumption and the tendency of the policy makers to introduce the tariff in order to promote energy efficiency.
As for the second aim of the analysis, panel specification appears appropriate both in terms of more significance of the estimates and of indexes of goodness of fit. Instead of the sub-optimal pooled specification, the random effects panel specification can be used as a better model especially when the phenomenon analyzed is observed in an heterogeneous contest. Furthermore, it can be a valid instrument to draw the individual specific features, especially when they are full of informative contents that, otherwise, could not completely emerge in a pooling specification. Finally, the purpose of the authors is to develop, based on empirical results obtained, methods able to estimate the probabilities of FiT adoption for countries that have not yet done it and to measure the effect of economic, environmental and generation factors on the decision to adopt FiT.

## Acknowledgements

The authors wish to thank the editor and two anonymous reviewers for detailed comments and suggestions. The usual disclaimer applies.

## References

[1] Akaike, H. (1974): A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716-723

[2] Amemiya T., McCurdy T.V (1986): Instrumental variable estimation of an error-components model. *Econometrica*, **54**, 869-950

[3] Carley, S. (2009) State renewable energy electricity policies: an empirical evaluation of effectiveness. *Energy Policy*, **37**, 3071-3081

[4] Chiung-Wen Hsu (2012): Using a system dynamics model to assess the effects of capital subsidies and feed-in tariffs on solar PV installations, *Applied Energy*, **100**, 205-217

[5] Croissant (2013): Pglm: panel generalized linear model. *R package version 0.1-2*, `http://CRAN.R-project.org/package=pglm`.

[6] Del Rio (2012): The dynamic efficiency of feed-in tariffs: The impact of different design elements. *Energy Policy*, **41**, 139-151

[7] De Luca, G., Carfora, A. (2014): Predicting U.S. recessions through a combination of probability forecasts. *Empirical Economics*, **46**, 127-144

[8] Dong C.G. (2012): Feed-in tariff vs. renewable portfolio standard: An empirical test of their relative effectiveness in promoting wind capacity development. *Energy Policy*, **42**, 476-485

[9] Estrella A., Mishkin F.S. (1998): Predicting U.S. Recessions: Financial Variables as Leading Indicators. *Review of Economics and Statistics*, **80(1)**, 45-61

[10] Greene W. (2003): Econometric Analysis fifth-edition. New York: Prentice Hall, 663-670

[11] Hausman, J., Taylor, W. (1981): Panel data and unobservable individual effects. *Econometrica*, **49**, 1377-1476

[12] IEA (2013): World energy outlook 2013. Tech. rep., International Energy Agency

[13] Islam, T., Meade, N. (2013): The impact of attribute preferences on adoption timing: The case of photo-voltaic (PV) solar cells for household electricity generation. *Energy Policy*, **55**, 521-530

[14] Kauppi, H., Saikkonen, P. (2008): Predicting U.S. Recessions with Dynamic Binary Response Models. *The Review of Economics and Statistics*, **90(4)**, 777-791

[15] Jakubson, G. (1991): Estimation and testing of the union wage effect using panel data. *The Review of Economic Studies*, **58**, 971-1062

[16] Lesser, J. A., Su, X. (2008): Design of an economically efficient feed-in tariff structure for renewable energy development. *Energy Policy*, **36**, 981-990

[17] Marques, A.C., Fuinhas, J.A., Pires Manso, J. R. (2010): Motivations driving renewable energy in European countries: a panel data approach. *Energy Policy*, **38**, 6877 - 6885

[18] McFadden,D. (1973): Conditional logit analysis of qualitative choice behavior. *In Frontiers in Econometrics (Edited by P. Zarembka)*, **105-42**

[19] Menz, F., Vachon , S. (2006): The role of social, political and economic interests in promoting state green electricity policies. *Environmental Science and Policy* **9**, 652-662

[20] Nyberg H., (2010): Dynamic Probit Models and Financial Variables in Recession Forecasting. *Journal of Forecasting*, **29**, 215-230

[21] Pinheiro J.C., Bates D.M. (1995): Approximations to the logLikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, **4(1)**, 12-35

[22] Romano A.A., Scandurra, G. (in press): 'Nuclear' And Non 'Nuclear' Countries: Divergences On Investment Decisions In Renewable Energy Sources. *Energy Sources Part B*, DOI:10.1080/15567249.2012.714843

[23] Romano A.A., Scandurra, G. (2011):The investments in renewable energy sources: do low carbon economies better invest in green technologies? *International Journal of Energy Economics and Policy*, **1(4)**, 107-115

[24] Romano A.A., Scandurra, G. (2014): Investments in Renewable Energy Sources in OPEC Members: a Dynamic Panel Approach. *Metodološki zvezki – Advances in Methodology and Statistics*, **11(2)**, 93 - 106

[25] Romano A.A., Scandurra, G., Carfora A (in press): Probabilities to adopt feed in tariff by OPEC countries using a panel probit model. *Energy Systems*, DOI: 10.1007/s12667-015-0173-5

[26] Romano A.A., Scandurra, G., Carfora A (2015): Probabilities to adopt feed in tariff conditioned to economic transition: A scenario analysis. *Renewable Energy*, **83**, 988 - 997

[27] Stokes, L. C. (2013): The politics of renewable energy policies: The case of feed-in tariffs in Ontario, Canada. *Energy Policy*, **56**, 490-500

[28] Toklu, E., Guney, M.S., Isik M., et al. (2010): Energy production, consumption, policies and recent developments in Turkey. *Renewable and Sustainable Energy Reviews*, **1**, 1172-1186

[29] UNEP (2012): *Feed in tariff as a Policy instruments for promoting Renewable energies and green economies in developing countries* `http://www.unep.org/pdf/UNEP_FIT_Report_2012F.pdf`

[30] Vella, F., Verbeke, M. (1998): Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics* **13**, 163-183

# Appendix

Estimation of model and all the data analyses were done using pglm package (Croissant, 2013) implemented in R statistical software. The package is available on the CRAN package repository (www.cran.r-project.org) while codes used to obtain reported results and all additional information useful to make research reproducible will be made available by the authors on request. Data employed are freely available from U.S. Energy Information Administration (www.eia.gov) and International Energy Agency (www.iea.org).

# INSTRUCTIONS TO AUTHORS

**Language**: *Metodološki zvezki – Advances in Methodology and Statistics* is published in English.

**Submission of papers**: Authors are requested to submit their articles (complete in all respects) to the Editor by e-mail (MZ@stat-d.si). Contributions are accepted on the understanding that the authors have obtained the necessary authority for publication. Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. Articles must be prepared in LaTeX or Word. Appropriate styles and example files can be downloaded from the Journal's web page (http://www.stat-d.si/mz/).

**Review procedure:** Manuscripts are reviewed by two referees. The editor reserves the right to reject any unsuitable manuscript without requesting an external review.

## Preparation of manuscripts

**Tables and figures**: Tables and figures must appear in the text (not at the end of the text). They are numbered in the following way: Table 1, Table 2,…, Figure 1, Figure 2,…

**References within the text:** The basic reference format is (Smith, 1999). To cite a specific page or pages use (Smith, 1999: 10-12). Use "et al." when citing a work by more than three authors (Smith et al., 1999). The letters a, b, c etc. should be used to distinguish different citations by the same author(s) in the same year (Smith, 1999a; Smith, 1999b).

**Notes:** Essential notes, or citations of unusual sources, should be indicated by superscript number in the text and corresponding text under line at the bottom of the same page.

**Equations:** Equations should be centered and labeled with two numbers separated by a dot enclosed by parentheses. The first number is the current section number and the second a sequential equation number within the section, e.g., (2.1)

**Author notes and acknowledgements:** Author notes identify authors by complete name, affiliation and his/her e-mail address. Acknowledgements may include information about financial support and other assistance in preparing the manuscript.

**Reference list:** All references cited in the text should be listed alphabetically and in full after the notes at the end of the article.

**References to books, part of books or proceedings:**

[1] Smith, J.B. (1999): *Title of the Book.* Place: Publisher.

[2] Smith, J.B. and White A.B. (2000): *Title of the Book.* Place: Publisher.

[3] Smith, J. (2001): Title of the chapter. In A.B. White (Ed): *Title of the Proceedings*, 14-39. Place: Publisher.

**Reference to journals:**

[4] Smith, J.B. (2002): Title of the article**.** *Name of Journal***, 2**, 46-76.

# Metodološki zvezki
## Advances in Methodology and Statistics


| | | |
|---|---|---|
| Published by | | Izdajatelj |
| **Faculty of Social Sciences University of Ljubljana, for Statistical Society of Slovenia** | | **Fakulteta za družbene vede Univerze v Ljubljani za Statistično društvo Slovenije** |
| Editors | **Valentina Hlebec**<br>**Lara Lusa** | Urednika |
| Founding Editors | **Anuška Ferligoj**<br>**Andrej Mrvar** | Prva urednika |
| Cover Design | **Bojan Senjur**<br>**Gregor Petrič** | Oblikovanje naslovnice |
| Typesetting | **Lara Lusa** | Računalniški prelom |
| Printing | **Littera Picta d.o.o.**<br>**Ljubljana, Slovenia** | Tisk |
| | **MZ** | |
| is indexed and abstracted in | **SCOPUS**<br>**EBSCO**<br>**ECONIS**<br>**STMA-Z**<br>**ProQuest** | je indeksirana in abstrahirana v |
| Home page URL | http://www.stat-d.si/mz/ | Spletna stran |

**ISSN 1854 - 0023**