ΙΔΡΥΜΑ ΟΙΚΟΝΟΜΙΚΩΝ & ΒΙΟΜΗΧΑΝΙΚΩΝ ΕΡΕΥΝΩΝ
FOUNDATION FOR ECONOMIC & INDUSTRIAL RESEARCH

Τσάμη Καρατάσου 11, 117 42 Αθήνα, Τηλ.: 210 92 11 200-10, Fax: 210 92 33 977, *www.iobe.gr*
11 Tsami Karatassou, 117 42 Athens, Greece, Tel.: +30 210-9211 200-10, Fax: +30210-9233 977

Weighting of responses in the Consumer Survey: Alternative approaches – Effects on variance and tracking performance of the Consumer Confidence Indicator

In the context of:

TASK FORCE ON THE QUALITY OF BCS DATA

Thomaidou Fotini
Valavanioti Evangelia
Vassileiadis Michail

November 2013

Table of Contents

1. Preface

The business and consumer surveys must continuously and accurately reflect the current situation of the underlying figures, their short-term tendency and their variations. In order to achieve that, the quality of their implementation must be ensured and assessed on a regular basis. To this end, certain qualitative criteria should be satisfied in each step of a business-consumer survey. These can take various forms {recommended procedures, minimum (statistical) requirements etc.}. On the other hand, it is probably not possible to produce strict norms for each step in a group of countries with different business and population characteristics or/and availability of census statistics, which limits statistical processing and testing. Nonetheless, some quality standards must be in power and their fulfillment can be monitored through various ways (survey metadata checks, statistical testing etc.).

The DG ECFIN Task Force on the Quality of the Business and Consumer Surveys (BCS) data focuses on the surveys' structural characteristics in all the partner countries (sampling frame, sample size, sampling method, response rate, weighting approach) and the parameters that affect them. These are linked to the quality of the DG ECFIN Business and Consumer Surveys data by the examination of their potential impact on volatility and tracking performance of the Business Climate and Consumer Confidence indicators. To this extent, the impact of the surveys' structural characteristics on the ability of these indicators to follow the trends of the underlying economic figures (sector production volume, consumption expenditure of households) and predict their changes in the short-term is assessed.

In this context, IOBE (EL) is the lead institute of the fifth thematic group for the Consumer Survey. STAT (FI), Gfk (PL) and IPSOS (HR) also participate in this thematic group. The examined topics by this thematic group concern the different weighting approaches used by the institutes participating to the DG ECFIN Consumer Survey, their structural features and the examination of their impact on volatility and tracking performance of the Consumer Confidence Indicator.

Our report is organised as follows. In the second section, a survey of the relevant literature is presented, in order to provide a thorough overview of the weighting approaches, in combination with the sampling techniques. The characteristics of the weighting approaches implemented in the Consumer Survey are presented in the third section, based on the institutes' answers to the questionnaire that IOBE has prepared, as well as on the available metadata and information from the DG ECFIN BCS website. In the fourth section, the impact of the different weighting approaches on volatility and tracking performance of the CCI is empirically investigated. Furthermore, linkages between certain weighting features and the variance of the CCI, as well as its correlation with the household-NPISH consumption are traced. Finally, in the fifth section the findings of the survey are summarised and some suggestions for adjustments on the weighting approaches that could reduce the CCI volatility are made.

2. A review of the theory on statistical survey weighting approaches

2.1 The most common types of weights in voluntary surveys

A sample must reflect the population it comes from and be representative with respect to all variables measured in a survey. However, this is not usually the case and non-response is one of the problems that can occur, resulting to some population groups being over- or under-represented in a sample. Another problem is self-selection that can appear in online surveys. If such problems occur, no reliable conclusions can be drawn from the survey data, unless something is done to correct for the lack of representativeness. The most commonly applied correction technique is weighting adjustment, a processing through the assignment of a certain weight to each survey respondent.

In the sense that was just described, the weights are the "corrective" values assigned to each one of the sample responses of a survey. In every data file, each case (response) normally has a weight. The weights are used primarily in order to make the computed statistics based on the gathered data more representative of the population from which the data are retrieved. Weights are often fractions, always positive and non zero. Individuals from under-represented population groups in a sample get a weight larger than 1, while those from over-represented groups get a weight smaller than 1. For example a weight of 2 means that the case counts as two identical cases in the data set. Then, in the computation of the means, totals and percentages, not just the values of the variables are being used, but the weighted values.

The most common types of weights are:[1]

i. **Design weights**
ii. **Post –stratification or non – response weights**
iii. **Population size weights**

These three methods are briefly explained below.

i. **Design weights**

The **design weights** are used: a) when we want the survey statistics to be representative of the underlying population or b) when we want to compensate for over- or under-sampling of specific cases or for disproportionate stratification. For computing design weights, we must know the sampling fraction, which is usually the over-sampling or the under-sampling amount for a given group or area. Thus, for instance, the unweighted samples in a survey over- or under-represent people of certain areas or size of households, such as those in larger households. The design weight corrects for differences in selection probabilities, thereby making the sample more representative of a 'true' sample of individuals in a country. The design weights are computed as normalised inverse of the inclusion probabilities. That is, if we know

[1] Johnson, D.R. (2008). "Using Weights in the Analysis of Survey Data". Population Research Institute

the sampling fraction of each respondent to the survey, then the weight is the inverse of the sampling fraction.

ii. Non-response Weighting

In voluntary surveys, one of the major threats for the accuracy of the estimates is non-responsiveness by the survey units. Different surveys achieve different response rates and the surveys that have low or declining response rate might suffer from severe survey bias in case this is not properly treated. However, **non-response is a problem if the non-respondents are a non-random sample of the total sample**, which is usually the case.[2] In household surveys, for instance, there is lots of evidence that non-respondents come from the younger strata of the population. In addition, it is relatively more difficult to persuade men to take part to surveys than women. As a result, the achieved survey samples often do not accurately reflect the underlying population and they may over-represent or under-represent some of its portions.[3] Accordingly it is common to put weights to sample survey datasets in order to compensate for this bias. This is known as "**non-response weighting"**. In this scope, non-response weighting is used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey and for this reason it is used for **handling unit non-response in surveys**.

The implementation of non-response weighting is usually a more complex process than design weights. Post stratification weighting, an alternative, commonly used designation for non-response weighting, usually requires further information about the underlying population of the sample survey and urges taking a number of different variables into account. There are many respondent characteristics (auxiliary statistics) that are likely to be related to the propensity to respond. The information usually needed is the population estimates of the distribution of a set of demographic characteristics that are inherent in the selected sample. Relative Information found in the Population Census is usually required, such as age distribution, educational level distribution, household size, race/ethnicity, gender, residence (e.g., rural, urban, metropolitan), region distributions etc. Accordingly, these distributions are compared with the factors/ variables distributions in the sample, stemming from the completed interviews, in order to proceed to the necessary adjustments that ensure representativeness.

iii. Population size weights

There is also a third weighting technique, the **population size weighting**. This is used when examining a combination of survey data from two or more countries –such as from international business and consumer surveys - and it corrects for the fact that most countries taking part have very similar sample sizes, no matter how large or small their population is. In this case, **the data must be adjusted in order to reflect the population size of each country**. Without this kind of weighting, any figures representing two or more countries would be inconsistent with the population they represent, resulting to over-representativeness of smaller countries. The population size

[2] Economic and Social Research Council. "Adjusting for non-response by weighting"
[3] In sample surveys women are usually over-represented and those over the age of 30. Furthermore, people living in cities and deprived areas are often under-represented.

weighting enables adjustment to ensure that each country is represented in proportion to its population size.[4]

2.2 Weighting design and most commonly applied weighting approaches

Most non-response weighting schemes involve **"post-stratification"** as it has already been mentioned, which is in essence a two-step procedure:[5]

(i) Identify a set of "control totals" of the population that the survey ought to match;

(ii) Calculate weights to adjust the sample totals to the control totals[6]

In its simplest form, post-stratification compares an N-way table from the population with an equivalent N-way table from the sample. A weight is calculated per cell of the table to adjust each observation to the population. In statistical analysis, **only one weight per case can be used**. If we weight for different factors, then these weights must be jointly taken into account in one weight for each case. The value of 0 cannot be a weight value, unless a specific case is excluded from the analysis. The default weight is equal to 1. A simple example is given below.

Table 2.1: A weighting example

Gender	Population Proportion	Sample Proportion	Population/Sample	Weight
Female	0.3	0.7	0.3/0.7	**0.428**
Male	0.5	0.4	0.5 /0.4	**1.25**
Total	1	1		

When several characteristics are jointly balanced, it is better to use several separate frequency tables rather than one big N-way crosstab to compute weights. There are different options to compute the weights and this creates various problems in the weighting procedure, since the different methods alter the weights and ultimately the impact of weighting.[7]

1. **Multiplication of the various weights**: This method involves the computation of a weight for each population characteristic independently and then the multiplication of the weights for all the characteristics that are taken into account. This method usually does not produce accurate weight estimates.
2. **Sequential computation of the weights**: This method involves the computation of the various weights separately, but sequentially. The first factor weights are calculated and then the population and sample distributions are compared. Then, we weight the sample data by the first factor weight. Next, we generate the frequency distribution for the second factor, after the data have been weighted by the first factor. We calculate the second factor weights. We weight the data by the first and second factor (by multiplying with the

[4] European Social Survey (2006). "Weighting European Social Survey Data". Norwegian Social Science Data Services (NSD)
[5] Another approach to treat non-responsiveness in surveys is imputation, but its presentation goes beyond the scope of the current survey.
[6] Economic and Social Research Council. "Adjusting for non-response by weighting"
[7] Ibid.

weights) and generate the weighted third factor frequency distribution. Then we calculate the third factor weight and so on.

This second approach is considered better, but the computed adjustment with respect to the population characteristics early in this sequence is not likely to match the underlying population structure when the characteristics taken subsequently into account are adjusted. This problem can occur **when the characteristics are correlated** (e.g. age and education).

The main constraint on post-stratification is that we need to know the exact population distributions. This automatically limits the control totals that can be used, which can be only the ones that are available from respective statistics and thus accurate. Thus, in the majority of national surveys, control totals tend to be age and sex within geographical areas. Statistics on other control totals that are considered useful in producing accurate estimates from samples are usually not available (e.g. social class).

There are alternative methods, besides simple post-stratification,[8] to use in the setting where the full N-way table for the population is not available, but the marginal distributions are.[9] These include:

a. **Raking or Iterative Solutions / proportional fitting:** Manual version (stepwise programming in statistical software) and Automatic version (Raking software, which is relatively widely used, usually in SAS and Stata statistical software).
b. **Logistic regression** based solutions, in case level population data is available. According to the literature, the regression models effectively smooth the weights so as to get more stable estimates, but usually yield weights that are highly correlated with those obtained by raking. Regression models can also be used in simpler circumstances as an alternative to simple post-stratification. They are mostly useful when a lot of information is available from the population and might result in weights with high variation that have high sampling errors.
c. **Calibration weighting**, which has become very popular for surveys of individuals selected via households, especially in the case where more than one individual is selected per household. The reason is that standard post-stratification will tend to give a different non-response weight to each member of a household and this can create difficulties for household-level analyses. Calibration uses an iterative procedure to create weights that bring individual level survey data into line with the population, but with the constraint that all individuals within a household must have the same weight. The underlying assumption is that non-response is primarily a household decision rather than an individual decision and it is household level non-response that creates most of the discrepancy between the sample and population distributions.

[8] Which are sometimes referred as sub-methods of post-stratification
[9] If for example, the gender distribution and the age distribution are known, but not the sex by age distribution

2.3 Differences in non-response and design weights

Not everybody has the same probability of selection in a survey. In probability sampling, each person's probability is known. When every element of the population has the same probability of selection, this is known as an **"equal probability of selection" (EPS) design**. Such designs are also known as **"self-weighting" samples** because all sampled units are given the same weight.

The main difference between design and non-response weights is the fact that the former can usually be accurately computed, but the latter are only estimated. In design weights, we know how many units are selected and how many were in the sampling frame.

In probability – based sampling, non-response weights are estimated by comparing the responding units to totals from the population or from the sampling frame. To produce results, we combine the responses from the sample in a way that takes into account the selection probabilities. If the sampling procedure were repeated many times we would get different numbers of non-responding units in each post survey stratum. This would give different non-response weights in each possible sample. **This uncertainty in the exact value of non-response weights should be reflected in the standard errors of the non-response adjusted analyses**. If the sampling procedure were to be repeated infinitely, the expected value of the results from the sequential samples would tend to be the same as the result we would get if we surveyed the whole population. Because we know the probability of getting each sample we select, we can also calculate a sampling error for the results. The sampling error tells us the amount of variation in the results due to the sampling alone, by providing at the same time a measure of the quality of the sample design and results.

The **design weights** are used in **non-probability sampling** (or purposive selection or judgmental selection), which is any sampling method where some elements of the population are not taken into consideration during selection (these are sometimes referred to as "out of coverage"/"undercovered"), or where the probability of selection cannot be accurately determined. This procedure involves the selection of population units based on assumptions regarding the population of interest, which forms the selection criteria. Hence, **because the selection of units is nonrandom, non-probability sampling does not allow the estimation of sampling errors.** These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. As information about the relationship between the sample and population is limited, it is difficult to extrapolate the sample results to population level.[10]

One of the most common methods of non-probability sampling is quota sampling. In quota sampling the selection of the sample is non-random and the population is first segmented into mutually exclusive subgroups. Thus, **quota sampling resembles the method of probability-based stratified sampling.** It is different however, because although the interviewers are constrained by the quotas, they are using their judgment in the choice of the sampled units. Since personal judgment is used to select the survey units from each segment, related to a specific population fraction, it

[10] Doherty, M. (1994). "Probability versus Non-Probability Sampling in Sample Surveys", *The New Zealand Statistics Review* March 1994 issue, pp 21-28.

is exactly this step that makes the technique a non-probability sampling one. **In general, non-responsiveness in a quota sample is handled by the selection of another respondent fitting the quota**. The problem is that the sample may be biased because not everyone has the same probability of selection. In other words, the lack of randomness is its biggest drawback. Quota sampling versus probability sampling has been accompanied by a great bulk of controversy. Moreover, in practice the application of the quota method is not always very clear as to what exactly is being done during the extraction of the sample from the population, in the sense that a quota sample may be drawn in stages, with the earlier stages involving selection (e.g. geographic areas) following probability-based methods, with quota sampling techniques used only in the last stages of sampling.[11] The quota sampling method usually requires availability of precise data over the whole population in order to set the quotas without bias.

2.4 Effects of non –response weighting on standard errors

The main motivation behind weighting for non-response is to remove bias and adjust the sample survey means and proportions to population standards. In specific, the method of post-stratification, relative to only applying inverse selection weights as it is usually in the case of the design weights, should, if done correctly, reduce the standard errors of most survey estimates, with exceptions being: a) the case where the variables used as control totals are unrelated to the survey variables and b) the case where there are small numbers of extremely large weights.[12] Usually, after the calculation of the weights, a check must be done for both these possibilities. It is a good practice to check the distribution of the weights, and if there are some very large weights, understand how and why they have arisen, in order to correct for errors. One standard practice to correct for (b) above is to trim very large weights. However, trimming the weights can also result in reducing the representativeness of the weighted data. Moreover, post-stratification is followed by other problems as well, which are related to the fact that a) it relies on the totals being correct and b) if post-stratification cuts across strata, correcting for standard errors requires more complicated methods of analysis, such as replication or calibration approaches. Therefore, **one problem of weighting is that, although it is a good tool for descriptive data, if not used properly, it may adversely affect inferential data and standard errors.** Weights, especially very large or very small ones, can also introduce instabilities into the data and increase the standard errors of the sample estimates.

On the other hand, the self-weighted datasets are often not efficient and can have lower statistical power than the weighted ones. Therefore, **there seems to be a tradeoff between less instability and less standard errors or more accurate representativeness.** However, if the sample is not self-weighted, then it is better to use weights as often as possible.[13]

Conclusively, **non-response weighting is a commonly used method for handling unit non-response in surveys and reducing non-response bias, but it is sometimes, under certain conditions, accompanied by a standard error increase**. However, **this is not**

[11] Doherty, M. (1994). "Probability versus Non-Probability Sampling in Sample Surveys", *The New Zealand Statistics Review* March 1994 issue, pp 21-28.
[12] Economic and Social Research Council. "Adjusting for non-response by weighting"
[13] European Social Survey (2006). "Weighting European Social Survey Data". Norwegian Social Science Data Services (NSD)

always the case, since non-response weighting can in fact lead to a reduction of variance as well as to lower bias. A covariate for a weighting adjustment must have two prerequisites to reduce non-response bias: a) be related to the probability of response, and b) be related to the survey outcome. If the latter is true, then weighting can reduce, not increase, sampling variance.[14]

3. Weighting approaches followed in the DG ECFIN Consumer Survey

In this section, the different weighting approaches followed by the institutes participating in the DG ECFIN Consumer Survey are presented, as well as their main defining characteristics. The relevant information was extracted from:

a) The questionnaire that IOBE sent to the institutes in the context of the thematic group of the BCS Task Force on weighting approaches used in the DG ECFIN Consumer Survey.

b) The Consumer Survey metadata available at the relevant DG ECFIN webpage.

The following tables summarise the information extracted from the two aforementioned sources. The response rate of the questionnaires, sent by IOBE to all the institutes participating in the DG ECFIN Consumer Survey, was rather high (24 out of 30 countries replied).[15]

Regarding the weighting approaches used by the institutes participating in the Consumer Survey, the majority of them uses a **random sampling method** (simple, stratified or systematic) and only a few use **quota sampling** (table 3.1). In specific, 10 and 19 out of 31 countries are using quota and random sampling respectively. The Consumer Survey in Spain and in Italy is conducted with the use of sampling methods which combine quota and random sampling.

The vast majority of the institutes apply a weighting method to the answers they get from the surveyed sample for the Consumer Survey. Almost all the countries that follow a probability sampling method put weights to the survey answers, with the exception of the Netherlands. Half of the countries implementing quota sampling apply design weighting, although weighting is not compulsory in non-probability sampling, provided that the sample drawn from the population is "nationally representative". Regarding the two countries that apply a combination of quota and random sampling, Spain uses weights whereas Italy does not. Overall, 24 out of 31 countries use some form of weighting in the Consumer Survey (table 3.2).

[14] Little, R.J. and Vartivarian, S. (2005). "Does Weighting for Non-response Increase the Variance of Survey Means?". University of Michigan Working paper

[15] Countries participating in the DG ECFIN Consumer Survey from witch a filled-in questionnaire was received are: Belgium, Czech Republic, Germany, Denmark, Spain, France, Italy, Cyprus, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovenia, Slovakia, Finland, Sweden, Turkey, Bulgaria, Romania, Croatia, Montenegro, Estonia

Table 3.1: Sampling Methods followed by Institutes participating in the Consumer Survey

	Quota Sampling	*Random sampling (simple/stratified/other)*	*Other*
Belgium (BE)		✓	
Czech Republic (CZ)		✓	
Germany (DE)	✓		
Denmark (DK)		✓	
Greece (EL)		✓	
Spain (ES)			✓
France (FR)		✓	
Italy (IT)			✓
Cyprus (CY)		✓	
Lithuania (LT)		✓	
Luxembourg (LU)	✓		
Malta (MT)		✓	
Netherlands (NL)		✓	
Poland (PL)		✓	
Portugal (PT)		✓	
Slovenia (SI)		✓	
Slovakia (SK)	✓		
Finland (FI)		✓	
Sweden (SE)	✓		
Turkey (TR)	✓		
Bulgaria (BG)		✓	
Romania (RO)	✓		
Croatia (HR)		✓	
Montenegro (ME)		✓	
Austria (AT)		✓	
Estonia (EE)	✓		
Ireland (IE)	✓		
Latvia (LV)		✓	
Hungary (HU)	✓		
United Kingdom (UK)	✓		
Former Yugoslav Republic of Macedonia (MK)		✓	

Source: IOBE questionnaire / DG ECFIN Consumer Survey Metadata

The use of weights by institutes that implement quota sampling was cross-checked with the metadata of the DG ECFIN Consumer Survey, taking also into account the mentioned in their questionnaires weighting factors, in order to avoid the possibility of identification of the weighting elements with the quota design factors.

Table 3.2: Use of a weighting approach

	Yes	*No*
Belgium (BE)	✓	
Czech Republic (CZ)	✓	
Germany (DE)	✓	
Denmark (DK)	✓	
Greece (EL)	✓	
Spain (ES)	✓	
France (FR)	✓	
Italy (IT)		✓
Cyprus (CY)	✓	
Lithuania (LT)	✓	
Luxembourg (LU)	✓	
Malta (MT)	✓	
Netherlands (NL)		✓
Poland (PL)	✓	
Portugal (PT)	✓	
Slovenia (SI)	✓	
Slovakia (SK)		✓
Finland (FI)	✓	
Sweden (SE)		✓
Turkey (TR)		✓
Bulgaria (BG)	✓	
Romania (RO)	✓	
Croatia (HR)	✓	
Montenegro (ME)	✓	
Austria (AT)	✓	
Estonia (EE)		✓
Ireland (IE)	✓	
Latvia (LV)	✓	
Hungary (HU)		✓
United Kingdom (UK)	✓	
Former Yugoslav Republic of Macedonia (MK)	✓	

Source: IOBE questionnaire / DG ECFIN Consumer Survey Metadata

Regarding the reasons for applying weights, the most significant of them seems to be selection bias reduction, with almost half of the countries participating in the Consumer Survey (15 out of 31) including it among the reasons for weighting (table 3.3). Variance reduction is the second most common reason for weighting, as it is indicated by 9 countries. The change of the underlying population seems to be a less crucial factor, as it is referred as a reason for weighting by only 3 countries.

Table 3.3: Reasons for applying weights*

	Reduce selection bias	Reduce variance	Underlying population change	Other
Belgium (BE)	✓	✓		
Czech Republic (CZ)	✓			
Germany (DE)	✓	✓		
Denmark (DK)	✓	✓		
Greece (EL)	✓	✓		
Spain (ES)	✓			✓
France (FR)	✓	✓	✓	
Cyprus (CY)	✓			
Lithuania (LT)	✓	✓		✓
Luxembourg (LU)	✓			✓
Malta (MT)	✓			
Poland (PL)	✓			
Portugal (PT)		✓		
Slovenia (SI)	✓			
Finland (FI)		✓	✓	
Bulgaria (BG)	✓	✓		
Romania (RO)				✓
Croatia (HR)				✓
Montenegro (ME)	✓		✓	

*Information available only for countries that responded to the IOBE questionnaire
Source: IOBE questionnaire

Apart from the above reasons for weighting, 5 countries reported some other reasons such as:

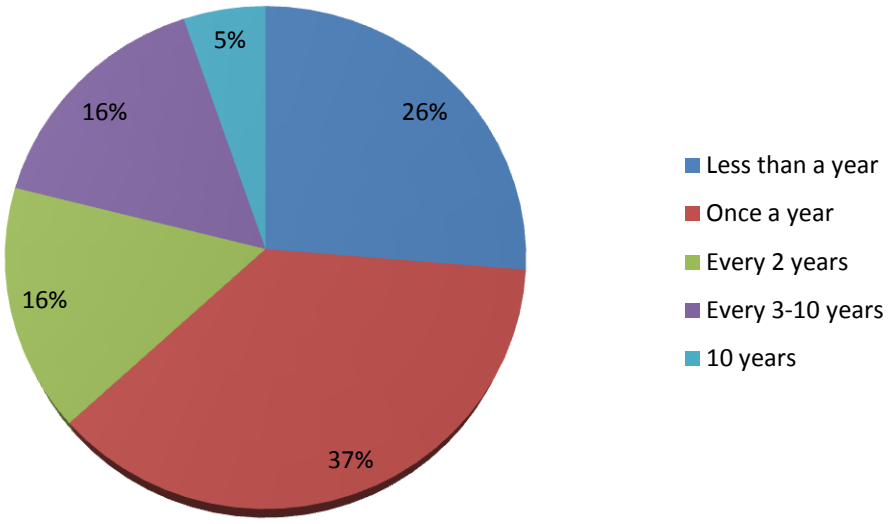
- ✓ Alignment of data with population figures
- ✓ Reduction of minimal differences of actual sample versus theoretical sample

However, these are also considered to be related to the reduction of selection bias and variance.

Differences among institutes do exist with regard to how often the weight coefficients are updated. Based on the answers to the IOBE questionnaire of 19 from the 24 countries that use weights, the weight update frequency varies from "every month" to "10 years" (figure 3.1). More than 1/3 of these countries (7 countries) update the weight coefficients once a year. 5 countries adjust them more frequently (each month or every quarter). Thus, the majority of the countries that follow a weighting approach (63%) update the weights at least once a year.

As it will be shown in a following subsection of this survey (4.4.1), the update frequency of weights is linked to high volatility in some of the countries that assign weights to the survey responses.

Figure 3.1: Update frequency of weights



Source: IOBE questionnaire

4. Empirical investigation of the impact of weighting approaches on the variance and tracking performance of the Consumer Confidence Indicator

4.1 Introduction

In this section, the existence of linkages between the weighting methods used by the institutes participating in the Consumer Survey and the volatility of the Consumer Confidence Indicator (CCI) is examined. The impact of different weighting approaches on the tracking performance of CCI, in specific of the fluctuations of households and NPISH consumption is also assessed.

Based on the alternative weighting approaches that were presented in the first section, as well as on the answers of the institutes to the questionnaire of IOBE, the institutes were categorised with respect to whether or not they weight the answers they get from the sample survey and the characteristics of their weighting approach. Hence, following the definitions given in the first section of this report, the determination of whether or not a country participating in the DG ECFIN Consumer Survey follows a weighting approach was linked to the implementation / non-implementation of non-response or design weighting. As it was mentioned in the first section of this survey, non-response weighting is usually required after probability-based sampling in order to correct for survey bias. Countries applying design weighting were also included in the country group with these that follow non-response weighting. Although design weighting is a process followed under a different sampling method, namely non-probability sampling (ex. quota sampling), many countries following non-probability sampling for the Consumer Survey do not

proceed afterwards to design weighting.[16] However, this is also a plausible processing of the sample responses, since in case where a sample drawn with non-probability sampling is considered "nationally representative", design weighting is not necessary. Accordingly, in only a few countries participating in the DG ECFIN Consumer Survey both non-probability sampling and design weighting are implemented (Germany, Ireland, Luxembourg, Romania, United Kingdom), making a group very difficult to handle for statistical inference. This is why they were categorised together with the countries that use non-response weighting. As it was already mentioned, the information needed for the distribution of institutes according to their weighting technique was also extracted by the country metadata of the DG ECFIN Consumer Survey available at the relevant DG ECFIN webpage.

Accordingly, in order to assess the impact of weighting on volatility and on the tracking performance of the CCI, the first country group includes countries that do not use a weighting approach and the second group countries that implement either non-response weighting or design weighting. The distribution of countries with respect to these criteria is shown in table 4.1. The majority of them, 24 out of 31, weight their answers. Five countries perform quota sampling and do not weight the collected answers. In Italy, a combination of probability sampling (random sampling) and non-probability sampling (quota sampling) is implemented, whereas in the Netherlands random sampling is used and answers are not weighted.

Table 4.1: Distribution of countries participating in the C.S. w.r.t. the weighting approach

No weighting	IT, EE, NL, SK, TR, HU, SE	23% of the DG ECFIN Consumer Survey countries
Weighting	DK, CY, MT, PL, SI, BG, HR, LU, BE, CZ, EL, LV, ES, FR, LT, PT, FI, AT, ME, MK, DE, RO, IE, UK	77% of the DG ECFIN Consumer Survey countries

Source: IOBE

With the help of the distribution of the institutes according to the weighting approach, their potential effect on the volatility of the Consumer Confidence Indicator (CCI) is tested in the following subsection. Next, the potential impact of the different weighting approaches to the tracking performance of the CCI is examined.

4.2 Weighting approaches and CCI volatility

The Months for Cyclical Dominance (MCD) index was adopted as the main measure of the volatility of the CCI in each country. The MCD index is a measure of short-term volatility for time series with monthly values. It is based on the decomposition of a time series to a trend-cycle component (C), a seasonal component (S) and an irregular component (I).[17] In case where a seasonally adjusted time series is decomposed, the seasonal component (S) has already been removed. The MCD index indicates the fewest number of months needed for the movement in the cycle component (C) to

[16] Estonia, Hungary, Slovakia, Sweden and Turkey do not proceed to design weighting or to any other weighting approach, although they implement non-probability sampling

[17] For a theoretical presentation of the decomposition steps see "Statistical Methods for Potential Output Estimation and Cycle Extraction (2003 Edition)", European Communities (2003). For details of the procedure in Eviews, see pp. 349 onwards of the EViews 7 User's Guide Vol.I.

dominate - on average over the examined period - changes in the irregular component (I) of a time series.[18] According to the OECD, there is a convention that the maximum value of the MCD index should be six months.[19]

A preliminary assessment of the volatility of the CCI was based on the ratio of the absolute change of its irregular component to the absolute change of its cyclical component in various time spans (one month, two months). Based on the estimations of this index and of the MCD index by DG ECFIN, a volatility analysis with respect to weighting approaches was carried out for 27 countries participating in the DG ECFIN Consumer Survey.[20] Estimations of the irregular to the cyclical component changes ratios and of the MCD index were made for two more countries, Turkey and Croatia, using the seasonally adjusted time series of the CCI available at the BCS time series webpage and the MCD computation methodology of DG ECFIN.[21]

The Mann-Whitney U test was used for assessing the statistical significance of differences in volatility between countries where a weighting method in the Consumer Survey is applied and countries that do not weight answers. The non-parametric nature of this statistical test implies relatively low limitations for the characteristics of the examined samples and the underlying populations.

As it has already been mentioned, a preliminary examination of the potential linkages between the weighting approaches and volatility was made using the estimations of the changes in the irregular to the changes in the cyclical component of the CCI ratio, for both the one-month time span $\{(I/C)_1\}$ and the two-month time span $\{(I/C)_2\}$. Taking into account the critical values of the MCD index set by the DG ECFIN in the presentation of the MCD estimations,[22] a convergence path for successive time spans of the value of the I/C index to below 1 was defined. According to that, values of the $(I/C)_1$ higher than 2.5, are considered indicative of high average changes of the irregular component of the CCI, relative to changes of the trend-cycle component and thus of high volatility of the CCI. For $(I/C)_1$ values lower than 2.5, the volatility of the CCI, as measured by the MCD index, is expected to be relatively low. In the same context, $(I/C)_2$ values higher than 1.5 are also considered indicative of the short-term dominance of the irregular component of the CCI over the trend-cycle component. On the contrary, in case where $(I/C)_2$ does not exceed 1.5, the underlying CCI time series is not expected to be volatile.

Based on the critical value of the I/C index for the one-month time span and the distribution of institutes with respect to the weighting approach they follow, the volatility of the Consumer Confidence Indicator is expected to be relatively low for the majority of countries that either apply a weighting approach or do not weight the sample survey responses: in almost 2/3 of the countries that weight their answers, the value of the $(I/C)_1$ is lower than 2.5. The respective proportion among countries that do not use a weighting approach is considerably higher, close to 85% (table 4.2).

[18] Gayer, C. (2010)
[19] "OECD Cyclical Analysis and Composite Indicators System - Users' Guide, Version 3" (2005)
[20] Data extracted from the Excel file "Quality indicators and metadata" sent by DG ECFIN in 18/02/13
[21] Described in the Word file "Months for Cyclical Dominance" sent by DG ECFIN on 27/03/13. The estimation of the MCD for Turkey and Croatia was based on the time series of the CCI, with values ranging from the starting period for each country until October 2012. No seasonally adjusted CCI data were found for FYROM and Montenegro.
[22] MCD value of 1 or 2: small volatility, MCD value of at least 4: high volatility. From the Excel file "Quality indicators and metadata"

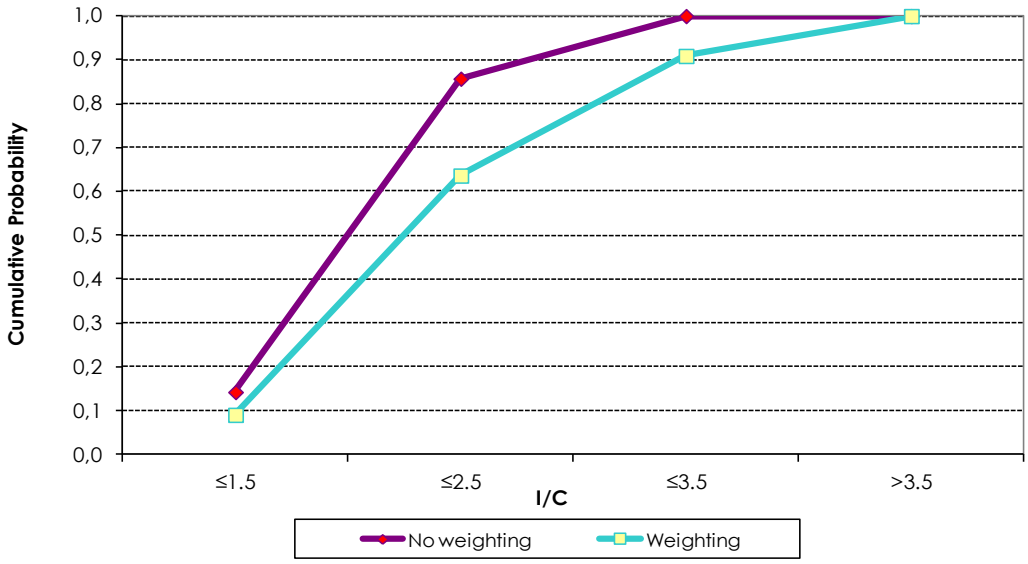
Table 4.2: Distribution of countries participating in the C.S. w.r.t. the weighting approach – magnitude of the I/C index (1-month time span)

	$(I/C)_1$	
	I/C>2.5	I/C ≤2.5
No weighting	IT	EE, HU, NL, SK, TR, SE
Weighting	DK, CY, MT, PL, SI, BG, HR, IE	BE, CZ, EL, LV, ES, FR, LT, PT, FI, AT, LU, DE, RO, UK

Source: IOBE

On the other hand, the difference in the central tendency of the $(I/C)_1$ index in the two country groups, as measured by its median value, is relatively small: the median of the $(I/C)_1$ among countries applying weights is 2.26, whereas for countries that do not use weights is 1.87.[23]

Figure 4.1: Cumulative distribution of $(I/C)_1$



Source: IOBE

The visual inspection of the cumulative distribution of the $(I/C)_1$ values for the two country groups reaffirms that some differences appear in the likelihood of its various levels between countries that apply weights and those that do not use weighting (figure 4.1). In 15% of the latter country group, the $(I/C)_1$ was lower than 1.5, a proportion close to that of the former group (11%). Differences between the two groups become more pronounced, when fractions of countries with higher $(I/C)_1$ values are taken into account. On the other hand, in 91% of the countries that use weights, the value of $(I/C)_1$ does not exceed 3.5, a proportion not significantly lower than that among the non-weighting countries (100%). **Thus, there are indications of lower volatility in those countries that do not weight their survey responses.**

Findings regarding the potential existence of a linkage between weighting approaches and volatility do not differ when taking into consideration the level of the (I/C) index for the two-month time span (table 4.3). The $(I/C)_2$ value does not exceed

[23] The median was preferred as a measure of central tendency from the average because, unlike the average, it is not affected by outliers and skewed data

1.5 in any of the non-weighting countries. On the contrary, in almost 1/3 of the countries that use a weighting approach, the value of $(I/C)_2$ is higher than 1.5. Nonetheless, the difference in the median of the $(I/C)_2$ index between the two country groups is smaller than it was in $(I/C)_1$: The median $(I/C)_2$ of countries applying weights is 1.29, whereas that of countries that do not use weights is 1.05.

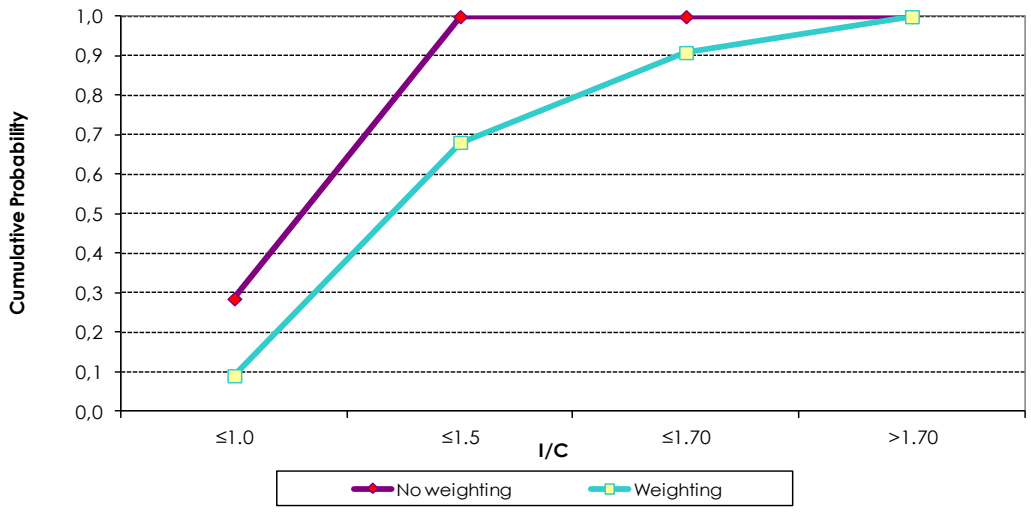
Table 4.3: Distribution of countries participating in the C.S. w.r.t. the weighting approach – magnitude of the I/C index (2-month time span)

	$(I/C)_2$	
	I/C>1.5	I/C≤1.5
No weighting		EE, IT, HU, NL, SK, TR, SE
Weighting	DK, CY, PL, SI, BG, HR, IE	BE, CZ, EL, LV, ES, FR, LT, PT, FI, MT, AT, LU, RO, DE, UK

Source: IOBE

The cumulative distribution of the $(I/C)_2$ for the two country groups shows that although $(I/C)_2$ exceeds 1.5 only in countries following a weighting approach, in the majority of these countries the level of this index is not significantly higher than 1.5 (figure 4.2). In specific, only in 9% of countries that weight the survey responses, $(I/C)_2$ exceeds 1.7.

Figure 4.2: Cumulative distribution of $(I/C)_2$



Source: IOBE

Indications from the I/C index for different time-spans regarding the volatility of the Consumer Confidence Indicator in countries that weight the survey answers and countries that do not weight them do not significantly differ from the findings based on the MCD index, which is adopted as the measure of volatility in our survey. Taking into account the critical values of the MCD index set by DG ECFIN, the volatility performance of all the non-weighting countries and of almost 65% of the countries that apply weights falls within the area of MCD values where volatility is not assessed as high (MCD≤3 months, table 4.4).

Table 4.4: Distribution of countries participating in the C.S. w.r.t. the weighting approach – MCD level

	MCD		
	MCD= 1 or 2	MCD=3	MCD>3
No weighting	HU, SE	EE, IT, NL, SK, TR,	
Weighting	DE, LT	BE, EL, LV, ES, FR, PT, FI, MT, AT, LU, RO, UK	DK, CY, PL, SI, BG, HR, CZ, IE

Source: IOBE

Further refinement of the volatility performances of institutes participating in the DG ECFIN Consumer Survey using the MCD index, reveals that their highest concentration is observed in MCD levels where volatility cannot be assessed as either low or high: in 72% of the non-weighting countries and 55% of the weighting countries, the value of the MCD index was estimated to be higher than 2 months and not greater than 3 months. Accordingly, the proportion of countries with a low MCD value - implying that the cycle component (C) of the CCI dominates in the very short run its irregular component (I) - is three times higher among the former country group than in the latter (28% vs. 9%).

The distribution of countries with respect to their MCD level provides some evidence of lower volatility among countries where the institutes do not apply weights. This fact does not imply that the volatility difference among the two country groups is statistically significant, since no relevant statistical measures have been used up to now in our analysis. Therefore, we next proceed to the statistical evaluation of the volatility of the CCI, both in the weighting and in the non-weighting countries, as well as of their difference.

The median of the MCD is the same in both country groups. Its magnitude is three months. Thus, in none of the two groups the central tendency of the volatility can be assessed as low.

Regarding the testing of the statistical significance of the difference of the volatility between the two country groups, a sample normality test was first applied, in order to test the null hypothesis (H_0) that the volatility data are drawn from a normally distributed population against the alternative hypothesis (H_1) that the underlying population is not normally distributed:

H_0: The sampled population is normally distributed

H_1: The sampled population is not normally distributed

In table 4.5, the result of the Shapiro-Wilk normality test is presented. The Shapiro-Wilk test was selected among other normality tests since it is more appropriate for small sample sizes (<50 observations) and it can also be applied to samples as large as 2000 observations. Since the P-value of the Shapiro-Wilk test for the MCD in countries participating to the Consumer Survey is less than 0.05, the null hypothesis of an underlying normally distributed population is rejected.[24]

[24] All the statistical tests were made with SPSS 16.0

Table 4.5: Normality test of the MCD distribution in countries participating to the C.S.

	Shapiro-Wilk test		
	Test function Value	df	H_0 Significance (P-value)
MCD	.825	29	.000

Source: IOBE

Due to the uncertainty about the distribution of the MCD over the sampled population, a non parametric test about the statistical significance of the difference of the volatility between countries that weight responses and countries that do not, according to the respective classification in table 4.1, was used. In specific, the Mann-Whitney U statistical test was applied. Besides the fact that it is a non parametric test, its selection was also based on the - not very restrictive- conditions that should be satisfied in order for the result of the test to be valid. In specific, the following assumptions should hold:

1. All the observations from both groups are independent of each other
2. The events are ordinal
3. There is symmetry between populations with respect to the probability of random drawing of a larger observation

The aforementioned conditions are theoretically satisfied by the MCD index, for the two country groups. No evidence exist that the magnitude of the MCD index for the CCI in a country is dependent of its magnitude in one or more countries. It is evident if the value of the MCD index in a country is greater or smaller than that in another country, since the MCD index takes only discrete values. In addition, no evidence exist that the probability of an MCD observation from the group with the weighting countries exceeding an MCD observation from the other country group, does not equal the probability of an observation from the latter group exceeding an observation from the former.

The null hypothesis (H_0) of the Mann-Whitney U test in our case, is that the volatility of the two underlying populations, as measured by the MCD index, is the same and it is tested against the alternative hypothesis (H_1) that in one of them the volatility tends to be higher.

H_0: distribution of MCD_1= distribution of MCD_2,

H_1: distribution of $MCD_1 \neq$ distribution of MCD_2,

where 1 and 2 refer to the two country groups

Given that the P-value of the Mann-Whitney U test function for the MCD level in the two country groups is marginally lower than 0.05, the null hypothesis is rejected at the 5% level of statistical significance (table 4.6). Thus, the difference between their volatilities is statistically significant. Taking into account the mean rank in each group, **the volatility of countries that do not use weights is lower than that of countries applying a weighting approach** (table 4.7).

Table 4.6: Mann-Whitney U test for the difference of the volatility between non-weighting & weighting countries

Mann-Whitney U	42.000
Z (test function value)	-2.016
H_0 Asymp. Significance (2-tailed)	.044
H_0 Exact Significance [2*(1-tailed Sig.)]	.078[a]

a. Not corrected for ties
Source: IOBE

Table 4.7: Mean rank of the non-weighting / weighting countries in the Mann-Whitney U test for the difference in volatilities

Group	N	Mean Rank	Sum of Ranks
Non-weighting	7	10.00	70.00
Weighting	22	16.59	365.00
Total	29		

Source: IOBE

To summarise the findings regarding the volatility of the Consumer Confidence Indicator with respect to the use of weights, some evidence of lower volatility in countries that do not apply weights were found from the assessment of the MCD level in the two country groups with respect to their weighting approach.

The Mann-Whitney U test result showed that the difference in the volatilities of the two country groups was statistically significant. It also showed that the variance of the CCI is lower in countries that do not assign weights to the Consumer Survey responses.

Although the lower volatility of the CCI in countries that do not do not apply any weighting approach was highlighted and verified through various ways (indications from the I/C index, MCD level, Mann-Whitney U test), it must be treated with cautiousness, since the sample of the non-weighting countries was very small (only seven countries) and thus not sufficient for statistical inference.

Nonetheless, this result does not contradict the theoretical framework presented in the first section of our survey regarding the non-response weighting, according to which *"non-response weighting in surveys… is sometimes, under certain conditions, accompanied by a standard error increase. These conditions could involve: a) the use of variables as control totals that are unrelated to the survey variables and b) the existence of a small number of extremely large weights."*

The potential sources of higher variance in the countries that apply weights are examined in a subsequent subsection (4.4). In specific, potential linkages between certain features of the weighting procedure and the volatility of the CCI are assessed. Their effects are evaluated together with these on the tracking performance of the CCI. Other potential sources of higher volatility, such as the existence of a small number of extremely large weights would require access to the primary survey data and weights of the institutes for a long period of time and are thus not feasible to be examined

We next move to tracing potential linkages between weighting approaches and the tracking performance of the Consumer Confidence Indicator.

4.3 Weighting approaches and CCI tracking performance

In order to evaluate the tracking performance of the Consumer Confidence Indicator, the DG ECFIN estimations of the correlation coefficient between the monthly level of the CCI and the corresponding year on year percentage change of seasonally adjusted household and NPISH consumption for 27 countries, were used.[25]

In order to define the critical level of correlation, above which the tracking performance of the CCI is considered relatively good, the relevant evaluation criteria of DG ECFIN in the presentation of the correlation coefficient estimations were taken into account.[26] Nonetheless, as the evaluation of the tracking performance was based not only on the contemporary correlation between the CCI and household-NPISH consumption, but also on the predictability of the Consumer Confidence Indicator of short-term GDP trends (i.e. the correlation of the indicator with future consumption fluctuations), the critical correlation coefficient level was different in each case. Since the CCI is mainly defined by expectations about developments in the following quarter, it should better reflect consumption fluctuations in the near future. Accordingly, stricter criteria in the evaluation of its tracking performance were used for its correlation with the consumption changes two and three months ahead. Correlation is expected to decline afterwards.

In specific, regarding the contemporary correlation, the critical value for the correlation coefficient was set at 0.60. In the majority of countries not using weights, the value of the correlation coefficient of the CCI with consumption changes was higher than 0.6 (table 4.8). On the contrary, in almost 2/3 of the countries applying a weighting approach, the estimated value of the contemporary correlation coefficient was smaller.

Table 4.8: Distribution of countries participating in the C.S. w.r.t. the weighting approach – correlation with household-NPISH consumption (contemporary)

	CONTEMPORARY CORRELATION	
	CORRELATION>0.60	CORRELATION≤0.60
No weighting	IT, HU, NL, SE	EE, SK
Weighting	EL, ES, LV, LT, PT, BG, RO, IE	BE, DK, CY, MT, FI, SI, AT, PL, FR, CZ, LU, DE, UK

Source: IOBE

The tracking performance in terms of correlation with the household consumption fluctuations of the next period of non-weighting and weighting countries is the same as in the contemporary correlation case (table 4.9). It is stressed out that there are no variations, even in the distribution of countries among the different performance categories, despite the fact that a higher critical value was set for the correlation coefficient (0.70).

[25] Files "Quality indicators and metadata" and "all graphs" sent by DG ECFIN on 18/02/13
[26] Excel file "Quality indicators and metadata"

Table 4.9: Distribution of countries participating in the C.S. w.r.t. the weighting approach – correlation with household-NPISH consumption (1 period ahead)

	CORRELATION (1 PERIOD AHEAD)	
	CORRELATION>0.70	CORRELATION≤0.70
No weighting	HU, NL, IT, SE	SK, EE
Weighting	LT, ES, EL, LV, PT, BG, RO, IE	FR, FI, SI, DK, MT, CY, AT, BE, PL, CZ, LU, DE, UK

Source: IOBE

Changes in the tracking performance of the weighting approaches were observed relative to the tracking household-NPISH consumption fluctuations two periods ahead: With the critical correlation coefficient value set to 0.70, more than half of the countries not using weights that had achieved a good tracking performance of the contemporary and one period ahead private consumption changes, failed to do so for a two-month time interval (table4.10). Nonetheless, the tracking performance of countries using weights was totally unaffected. Consequently, their average tracking performance of the private consumption fluctuations exceeds that of the countries not using weights.

Table 4.10: Distribution of countries participating in the C.S. w.r.t. the weighting approach – correlation with household-NPISH consumption (2 periods ahead)

	CORRELATION (2 PERIODS AHEAD)	
	CORRELATION>0.70	CORRELATION≤0.70
No weighting	HU	SK, EE, SE, NL, IT
Weighting	LT, ES, EL, LV, PT, BG, RO, IE	FR, FI, SI, DK, MT, CY, AT, BE, CZ, PL, LU, DE, UK

Source: IOBE

The tracking performance of the Consumer Confidence Indicator is not negatively affected if the time interval between the reference period of the CCI and the period ahead for which the household-NPISH consumption change is tracked is increased to three months. With the correlation coefficient critical value set to 0.65, two more countries, one among these applying a weighting approach and the other among these that do not, achieve a good tracking performance (the Netherlands and Czech Republic respectively, table 4.11). Accordingly, **the proportion of countries achieving a good tracking performance remains higher among countries that use weights but its difference with the respective proportion in the non-weighting countries is smaller than it is as regards tracking performance two periods ahead** (43% and 33% respectively).

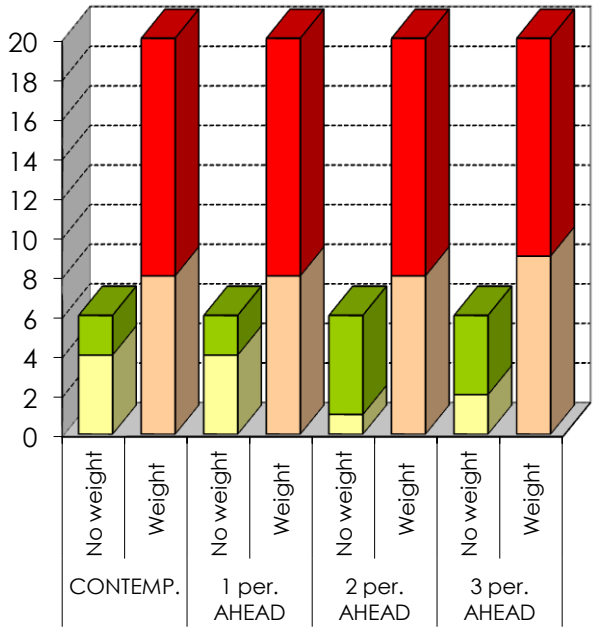
Table 4.11: Distribution of countries participating in the C.S. w.r.t. the weighting approach – correlation with household-NPISH consumption (3 periods ahead)

	CORRELATION (3 PERIODS AHEAD)	
	CORRELATION>0.65	CORRELATION≤0.65
No weighting	HU, NL	SK, EE, SE, IT
Weighting	LT, ES, EL, LV, PT, BG, CZ, RO, IE	FR, FI, SI, DK, MT, CY, AT, BE, PL, LU, DE, UK

Source: IOBE

To sum up the findings for the tracking performance of the CCI with respect to the different weighting approaches, the countries that do not weight the survey responses perform better than these that assign weights, in terms of contemporary and one period ahead correlation with the fluctuations of the household-NPISH consumption expenditure. On the other hand, when the correlation of the CCI with changes in private consumption two or three months ahead is examined, its tracking performance in countries that apply weights is better than that of countries that do not use weights. All these findings are depicted in figure 4.3. Starting from the contemporary correlation results to the left side, the proportion of the weighting countries that achieve a good tracking performance remains unaffected by the increase of the time interval between the reference period of the CCI and the period of interest regarding private consumption fluctuations. This proportion becomes higher for household-NPISH consumption three months ahead. On the contrary, after the three-month time interval, the proportion of the non-weighting countries with high correlation coefficient between the CCI and the private consumption falls. As a consequence, it becomes lower than that in the countries that apply weights.

Figure 4.3: Trends in the tracking performance of the CCI w.r.t. weighting approach



Lower strata: Countries with high correlation
Upper Strata: Countries with low Correlation

Source: IOBE

Nevertheless, as in the case of the test on the effects of the weighting of responses to the volatility of the CCI, these findings must be interpreted with cautiousness, as the sample of the non-weighting countries is very small. Thus, a potential inclusion of a few more "observations" (countries) to it could drastically change their tracking performance.

4.4 Weighting features with a potential impact on the volatility and tracking performance of the Consumer Confidence Indicator

The volatility of the CCI (as measured by the MCD index) was found to be significantly lower in statistical terms in the countries that do not weight the sample responses from these that apply a weighting method. On the contrary, neither non-weighting nor weighting was found to improve the tracking performance of the CCI with respect to the fluctuations in household-NPISH consumption.

Thus, it is possible that some structural features of the weighting procedure are linked to higher CCI variance in countries that put weights to the sampled units according to certain population characteristics/strata. Accordingly, their proper adjustment could lead to more representative and reliable Consumer Survey results. After taking into account the topics of the other thematic groups of the Task Force on the Quality of BCS data in order to avoid overlapping of survey areas, we focus on the potential effects on the variance of the CCI of: a) the **population characteristics / strata included in non-response and design weighting** and b) the **update frequency of weights**. We also extent our analysis to the examination of the potential impact of these weighting features on the correlation of the CCI with the household-NPISH consumption.

It is straightforward that both these weighting features could have an effect only on those countries which use weights. They do not have an impact on non-weighting countries, thus comparisons between the two country groups cannot be made. On the other hand, in the non-weighting countries there are no features of a weighting approach to be examined. Consequently, in these countries, other Consumer Survey characteristics (e.g. sampling method, response rate, sampling frame) could be more clearly linked to potential weaknesses of the CCI in terms of volatility and tracking performance.

In the rest of this section, the impact of the two aforementioned weighting features on the MCD levels of the CCI is examined first. Next, we try to clarify their influence on its tracking performance of the household-NPISH consumption fluctuations.

4.4.1 Impact of weighting features on the volatility of the CCI

The cross tabulation of countries that apply weights and have a high MCD[27] together with the population characteristics they take into consideration in weighting, shows that no common pattern exists among them as to how many and which of these characteristics are included in the weights (table 4.12). As it usual in population surveys, **gender, age group and region of residence are the most commonly used**

[27] Bulgaria, Czech Republic, Croatia, Cyprus, Denmark, Poland, Slovenia, Ireland

population characteristics for the formation of the weights. Thus, it cannot be supported that in the case of countries which apply weights and have a high MCD, the relatively high volatility of the CCI is related either to certain population characteristics/ strata that are included in the weight coefficients or to the number of these strata.

Table 4.12: Population characteristics/strata affecting the weight coefficients (weighting countries with high volatility)*

Countries	Population Characteristics							
	Age	Gender	City size	Family size	Income	Education	Region	Other
Bulgaria	✓	✓	✓				✓	
Czech Republic	✓	✓	✓	✓		✓	✓	Economically active/inactive
Croatia		✓				✓	✓	
Cyprus	✓	✓						
Denmark	✓	✓		✓	✓	✓	✓	Ethnicity and dwelling type
Poland	✓	✓	✓	✓			✓	
Slovenia	✓	✓	✓			✓	✓	

* No data availability for Ireland
Source: IOBE

Regarding the update frequency of the weights, it is on average lower in countries with a high MCD. The average time interval between two weights updates is 3.7 years in countries with a high MCD, instead of 1.2 years in countries with lower MCD, whereas the median of the update frequency of the weights is 2.0 and 1.0 years respectively.[28] Furthermore, a Mann-Whitney U test on the update frequencies of the weights of the two country groups showed that at 10% level of significance these are not the same (table 4.13). As the mean rank of the countries with lower volatility does not exceed that of the countries with high MCD, the update frequency of the weights is significantly higher among the former country group (table 4.14).

Table 4.13: Mann-Whitney U test for the difference of the update frequency of the weights between high and low MCD weighting countries

Mann-Whitney U	21.000
Z (test function value)	-1.670
H_0 Asymp. Significance (2-tailed)	.095
H_0 Exact Significance [2*(1-tailed Sig.)]	.126[a]

a. Not corrected for ties
Source: IOBE

[28] No relevant data availability for Austria, Latvia and United Kingdom

Table 4.14: Mean rank of the weighting countries with high / low MCD in the Mann-Whitney U test for the difference of the update frequency of weights

MCD group	N	Mean Rank	Sum of Ranks
High MCD	7	12.l00	84.00
Lower MCD	11	7.91	87.00
Total	18		

Source: IOBE

In order to crosscheck this outcome and further extend the assessment of the effect of the update frequency of weights on the volatility of the CCI, the difference in its variance between countries using weights that update them at least once a year and countries that do not apply any weighting approach was examined. Applying again the Mann-Whitney U test, the H_0 hypothesis of same volatility in the two country subgroups was not rejected at the 10% level of statistical significance (table 4.15). It is stressed that in half of the countries that participate in the DG ECFIN Consumer Survey and assign weights to the sample responses (12 out of 24), these are updated once a year or more regularly.[29] Accordingly, the higher volatility of the CCI that was found in subsection 4.2 in countries using weights relative to countries that do not follow any of the weighting methods can be attributed to the high MCD level of these that do not update the weight coefficients at least once a year.

Table 4.15: Mann-Whitney U test for the difference of the volatility between countries with frequently updated weights & non-weighting countries

Mann-Whitney U	29.500
Z (test function value)	-1.229
H_0 Asymp. Significance (2-tailed)	.219
H_0 Exact Significance [2*(1-tailed Sig.)]	.299[a]

a. Not corrected for ties
Source: IOBE

Based on the above findings we can conclude that only the frequency of the weights updating tends to have a significant impact on the MCD level. If the weight coefficients are not frequently updated (i.e. at least twice a year=average frequency value in the countries frequently updating the weight coefficients) then the volatility of the CCI will tend to be increased. Conclusively, institutes that do not frequently update the weight coefficients of responses to the Consumer Survey would probably reduce the variance of their CCI estimations if they did so.

4.4.2 Impact of weighting features on the tracking performance of the CCI

Since a minority of the weighting countries achieved a good tracking performance of the changes in household-NPISH consumption by the CCI, we next examine if there are any similarities among them, in terms of the population characteristics they take into account in weighting, as well as in the update frequency of weights.

[29] Data availability for 19 countries using weights that responded to the relevant question of the IOBE questionnaire

Regarding the strata taken into consideration in the weights formation, almost all countries with a high CCI-private consumption correlation coefficient weigh their responses according to same characteristics, namely **gender, age group and region of residence** (table 4.16). The size of the city of residence is also widely used. Only Portugal assigns weights based on completely different population characteristics (family size and income). Nonetheless, the weights in the vast majority of these countries are determined by the most commonly used population characteristics/ strata.

Table 4.16: Population characteristics affecting the weight coefficients (weighting countries with good tracking performance)*

Countries	Population Characteristics							
	Age	Gender	City size	Family size	Income	Education	Region	Other
Bulgaria	✓	✓	✓				✓	
Greece	✓	✓	✓				✓	
Latvia	✓	✓					✓	Nationality
Lithuania	✓	✓						Living area
Portugal				✓	✓			
Romania	✓	✓	✓				✓	
Spain	✓	✓	✓				✓	

* No data availability for Ireland
Sources: BCS meadata / IOBE

The frequency of the weights updating is not different in countries with a good tracking performance relative to these with low correlation coefficients between the CCI and household-NPISH consumption. Its median is one year in both country subgroups.[30] A Mann-Whitney U test confirmed that the difference of the average update frequency between countries with high and relatively low correlation coefficients is not statistically significant, as the H_0 hypothesis can only be rejected at a very low level of significance (table 4.17).

Table 4.17: Mann-Whitney U test for the difference of the update frequency of weights between weighting countries with high and low CCI-household consumption correlation

Mann-Whitney U	30.500
Z (test function value)	-.267
H_0 Asymp. Significance (2-tailed)	.789
H_0 Exact Significance [2*(1-tailed Sig.)]	.808[a]

a. Not corrected for ties
Source: IOBE

Thus, we cannot assert that a good tracking performance of the fluctuations of the household-NPISH consumption by the Consumer Confidence Indicator is a result of certain population characteristics that were taken into consideration during weighting, or of a frequent update of the weight coefficients.

[30] No relevant data availability for Austria, Latvia and United Kingdom

5. Main Findings – Suggestions

The purpose of the thematic group on weighting approaches used in the Consumer Survey in the context of the Task Force on the quality of the BCS data was mainly to assess the impact of different weighting approaches used by the institutes participating to the survey on the volatility and the tracking performance of the Consumer Confidence Indicator.

The weighting approaches applied by the institutes participating to the DG ECFIN Consumer Survey are closely linked to the sampling method they use. For example, **design weighting is used in non-probability sampling** (ex. quota sampling), where the surveyed units are included to the sample according to previously made assumptions regarding the population of interest, which define the selection criteria. Accordingly, the selection of the surveyed units is nonrandom and can become a source of exclusion bias. In such cases, in order to make the survey statistics representative of the population, design weighting is applied. **In probability – based sampling** (ex. random sampling) the formation of the sample is not following some selection criteria. On the other hand, this fact creates space for non-responsiveness by certain population groups. If the non-respondents are a nonrandom sample of the total population, then the population estimations one would get from those that responded to the survey would be unbiased. In order to avoid such an event, **non-response weighting can be applied.**

The majority of the institutes participating to the DG ECFIN Consumer Survey **apply non-response weighting (19 out of 31 countries). Five countries implement design weighting and seven countries do not weight the sample responses**. **Although ten countries follow non-probability sampling, half of them do not use design weights**, which is also a plausible processing in case where the drawn sample is considered to be "nationally representative". Accordingly, only five countries implement both non-probability sampling and design weighting and are very difficult to handle as a group for statistical inference. This is why, in order to trace potential relationships between the different weighting methods (including no weighting) and volatility and tracking performance of the Consumer Confidence Indicator, countries were categorised with respect to whether they use any kind of weighting to the sample survey answers. Thus, countries applying design weighting and these that follow non-response weighting were included in the same country group.

The Months for Cyclical Dominance (MCD) index was adopted as the measure of the volatility of the Consumer Confidence Indicator. **Taking into account the critical values of the MCD for high / low volatility, we concluded that the countries that do not assign weights to the sample responses show lower CCI variance than those that apply non-response or design weights**. Given that the underlying population is not normal according to the result of the Shapiro – Wilk test, **the statistical significance of the difference in the volatilities of the two country groups was tested with the means of the Mann-Whitney U test**. The null hypothesis of same volatility among the two country groups was rejected. However, this is a **result that must be treated with cautiousness due to the very small sample of the non-weighting countries** (seven observations). If one more country with high volatility (MCD>3) was included to these that do not weight the survey responses, the result of the statistical test would have been different and the volatility of the two underlying populations would be the

same. Nonetheless, this result does not contradict the theoretical framework regarding non-response weighting, since it can be accompanied, under certain conditions, by a standard error increase.

Regarding the tracking performance of the CCI under the different weighting approaches, the **non-weighting countries perform better** than these that use weights **in terms of contemporary and one period ahead correlation** with the fluctuations of the household-NPISH consumption expenditure. On the other hand, when the correlation of the CCI with **changes in private consumption two or three months** ahead is examined, **the tracking performance of the CCI in weighting countries is better** than that of countries that do not use weights. Thus, the accuracy of short-term projections about the changes in the household-NPISH consumption based on the trend of the CCI is not better under either of the examined weighting approaches.

In order to track the sources of higher MCD among countries that use weights, the existence of a **relationship between certain weighting features and the volatility of the CCI was examined**. In specific, the **potential effects** of: **a) the population characteristics/strata included in non-response and design weighting and b) the update frequency of weights**, on the variance of the CCI were assessed. Other potential sources of higher volatility, such as the existence of a small number of extremely large weights were not feasible to be examined. We concluded that **no common pattern** exists **among the countries with high MCD as to how many and which of the population characteristics /strata are taken into account in the construction of the weights**. As it usual in population surveys, gender, age group and region of residence are the most commonly used strata. Thus, the **high volatility of the CCI** in these countries **is not owed to either the inclusion of certain strata characteristics to the weight coefficients or the number of the strata used**.

Concerning the impact on the MCD level of the second weighting feature under evaluation, the **difference of the update frequency of the weights between countries with high and low MCD level, was statistically significant** at the 0.10 level of significance and the median of the update frequency was lower at the second country subgroup. **Furthermore, when excluding the countries with a low update frequency of the weight coefficients from the country group of these that apply weights**, **the difference in its CCI volatility with the group of countries that do not follow any weighting approach is not statistically significant**. Thus, it can be supported that **the use of weights is not a source of higher volatility**, as it was found earlier, **provided that these are frequently updated (i.e. at least twice a year=** average frequency in the countries frequently updating the weights). Alternatively, a more regular update of the weights could contribute to lowering the variance of the CCI in the countries with a high MCD level participating to the Consumer Survey.

The **effect of the two weighting features on the tracking performance of the CCI was also examined**. **Almost all the countries with a high CCI-private consumption correlation coefficient weight their responses according to the same population characteristics** (gender, age group and region of residence), which are the most commonly used population characteristics in weighting. The countries with a good tracking performance do not take into consideration in weighting either some other qualitative characteristics of the population not used by the vast majority of countries that put weights, or a combination of the widely used parameters and of some

special population features. Thus, **the high correlation between the CCI and private consumption in some of the Consumer Survey countries that use weights cannot be attributed to a special designing of the weights w.r.t. the used population characteristics/strata**.

Finally, **no evidence of improvement of the tracking performance of the CCI by frequently updating the weight coefficients** was found in countries with high CCI-private consumption correlation coefficient.

6. References

Doherty, M. (1994). "Probability versus Non-Probability Sampling in Sample Surveys", *The New Zealand Statistics Review*, March 1994 issue, pp 21-28.

Economic and Social Research Council. "Adjusting for non-response by weighting") (http://www.esrc.ac.uk/)

European Communities (2003), Working Papers and Studies. "Statistical Methods for Potential Output Estimation and Cycle Extraction (2003 Edition)"

European Social Survey (2006). "Weighting European Social Survey Data". Norwegian Social Science Data Services (NSD)

Gayer, C. (2010). "Report: The Economic Climate Tracer. A Tool to Visualise the Cyclical Stance of the Economy Using Survey Data", EU Commission, Brussels

Johnson, D.R. (2008). "Using Weights in the Analysis of Survey Data". Population Research Institute

Little, R.J. and Vartivarian, S. (2005). "Does Weighting for Non-response Increase the Variance of Survey Means?". University of Michigan Working paper

OECD (2005). "OECD Cyclical Analysis and Composite Indicators System - Users' Guide, Version 3"

Quantitative Micro Software, LLC (2010). "EViews 7 User's Guide Vol.I"