

The Fourth Community Innovation Survey (CIS 4)

Methodological recommendations

(In accordance with section 7 of the annex to the Commission Regulation on innovation statistics No 1450/2004)

Final version 9 November 2004

0. Introduction

The Commission Regulation No 1450/2004, implementing Decision No 1608/2003/EC of the European Parliament and of the Council concerning the production and development of Community statistics on innovation (= Commission Regulation on innovation statistics), puts innovation statistics on a statutory basis and makes compulsory the delivery of certain variables. This document, which outlines the harmonized methodology to be used for CIS 4, is related to section 7, paragraph 2 of the annex of this Commission Regulation on innovation.

1. Target population

The target population of the CIS 4 shall be the total population of enterprises related to market activities (NACE activities C to K).

1.1. NACE

Core coverage

In accordance with section 2 of the annex of the Commission Regulation on innovation statistics, the following industries shall be included in the core target population of the CIS 4:

- mining and quarrying (NACE 10-14)
- manufacturing (NACE 15-37)
- electricity, gas and water supply (NACE 40-41)
- wholesale trade (NACE 51)
- transport, storage and communication (NACE 60-64)
- financial intermediation (NACE 65-67)
- computer and related activities (NACE 72)
- architectural and engineering activities (NACE 74.2)
- technical testing and analysis (NACE 74.3)

Additional coverage, in order of descending priority (to be done on a voluntary basis):

- research and development (NACE 73)
- construction (NACE 45)
- motor trade (NACE 50)

- retail trade (NACE 52)
- legal, accounting, market research, consultancy and management services (NACE 74.1)
- advertising (NACE 74.4)
- labour recruitment and provision of personnel (NACE 74.5)
- investigation and security activities (NACE 74.6)
- industrial cleaning services (NACE 74.7)
- miscellaneous business activities n.e.c. (NACE 74.8)
- real estate activities (NACE 70)
- hotels and restaurants (NACE 55)
- renting of machinery and equipment without an operator (NACE 71)

These economic activities should be regarded as “non-core” and do not necessarily have to meet the same quality requirements as for the core coverage e.g. for item and unit non-response (i.e. a non-response survey does not have to be carried out in respect of these NACE industries) or the required level of precision.

1.2 Size-classes

It is recommended that **all** enterprises be included in the target population. However, the minimum coverage shall be all enterprises with **10 employees or more**.

1.3. Statistical units

The main statistical unit for CIS 4 shall be the enterprise, as defined in the Council Regulation 696/1993 on statistical units or as defined in the national statistical business register. EU Regulation 2186/1993 requires that Member States set up and maintain a register of enterprises, as well as associated legal units and local units.

In the Council Regulation 696/1993¹, the enterprise is defined as “the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision making, especially for the allocation of its current resources. It may carry out one or more activities at one or more locations and it may be a combination of legal units, one legal unit or part of a legal unit.”

In general, innovation activities and decisions usually take place at the enterprise level, which leads to the enterprise being used as the statistical unit. If the use of the enterprise as a statistical unit is not feasible, other units such as the division of the enterprise group, the kind of activity unit (KAU), the local kind of activity unit (LKAU) or the enterprise group may be used instead.

1.4 The observation period

¹ Council Regulation (EEC) N° 696/1993 of 15 March 1993, OJ N° L76 of the 3 March on the statistical units for the observation and analysis of the production system in the Community.

The observation period to be covered by the survey shall be 2002-2004 inclusive i.e. the three-year period from the beginning of 2002 to the end of 2004. The reference period of the CIS 4 shall be the year 2004.

2. Survey methodology

2.1. Sampling frame

The **official, up-to-date, statistical business register² of the country** should be used.

2.2 Census or sample survey

Data should be collected through a census, sample survey or a combination of both.

2.3 Stratification

The target population shall be broken down into similar structured subgroups or strata (which should be as homogeneous as possible and form mutually exclusive groups). Appropriate stratification will normally give results with smaller sampling errors than a non-stratified sample of the same size and will make it possible to ensure that there are enough units in the respective domains³ to produce results of acceptable quality.

The stratification variables to be used for the CIS 4, i.e. the characteristics used to break down the sample into similarly structured groups, should be:

- The economic activities (in accordance with NACE)⁴.

In accordance with the requirements of section 5, paragraph 2 of the annex of the Commission Regulation on innovation statistics, stratification by NACE should be done at least at two-digit (division) level, except for NACE 74. Here the three digit sections NACE 74.2 and 74.3 should be treated as separate NACE categories while NACE 74.1 and 74.4 to 74.8 should be treated as a single NACE category.

- Enterprise size according to the number of employees⁵.

The size-classes used should at least be the following:

- 0-9 employees
- 10-49 employees
- 50-249 employees
- 250+ employees.

² Council Regulation (EEC) N° 2186/1993 of 22 July 1993.

³ Domains are defined as strata or combinations of strata, for which results will be published.

⁴ The NACE code to use for stratification should be that of the enterprise at the end of the reference period 2004.

⁵ The enterprise size to use for stratification should be the number of employees at the end of the reference period 2004.

More detailed breakdown by size classes may also be used, but, whatever size-classes are chosen, they should fit into the above size groups.

- Regional aspects:

In accordance with section 7, paragraph 2 of the annex of the Commission Regulation on innovation statistics, the methodology will include regional aspects. Therefore, the regional allocation of the sample shall be taken into consideration when sampling.

2.4. Sample size

There is no minimum sample size needed, as long as the sample size chosen will meet the precision levels required (see section 4.6). However, if a particular stratum has less than 6 enterprises, then all the enterprises in this stratum should be selected for the survey.

The expected response rate should be borne in mind i.e. the sample size should take into account the non-response rates experienced in CIS 3 and compensate accordingly. Finally, there should be no replacement of deleted or not-relevant units. The sample size should be large enough to compensate for any of these types of units.

2.5 Sample selection and allocation

The selection of the sample should be based on random sampling techniques, with known selection probabilities, applied to strata. It is recommended to use simple random sampling without replacement within each stratum.

Different allocation schemes can be used, depending on the structure of the population. It is recommended to use optimum allocation, taking into account the need to “compromise” the allocation, in order to obtain the required levels of precision for all indicators and domains.

The variance in each stratum to be used for sample selection can be based on previous CIS 3 results, if there is reliable information available. If not, one can either use the CIS 3 national average or assume that a problem stratum will be close to a stratum for which reliable results are available. If new sectors of the economy are added for the CIS 4, one can either use the national average for the CIS 3 or assume that the new sector will be close to a sector that has been sampled previously.

Member States are free to use whatever sampling methods they prefer, as long as the quality thresholds for the results are achieved. However, in accordance with section 7, paragraph 4 of the annex of the Commission Regulation on innovation statistics, Eurostat should be informed in advance of the method of sampling and allocation scheme being used.

3. Collecting and processing of data

3.1 SAS programs for processing the data

The SAS programs which were used for CIS 3 will be updated for use for the CIS 4 and provided free (along with good user documentation) to those Member States that want them⁶. There will be some user support for these programs once the CIS 4 starts. The program rules will also be provided.

3.2 Survey questionnaire

In accordance with section 7, paragraph 1 of the annex of the Commission Regulation on innovation statistics, the CIS 4 will be based on a harmonised survey questionnaire for all NACE sectors. The questionnaire shall cover the main themes listed in the Oslo Manual. This harmonised questionnaire shall be used in all national innovation surveys.

3.3 Data collection

The CIS 4, like the previous innovation surveys, shall be mainly based on mail surveys. These provide a relatively inexpensive means of gathering information from a widely dispersed sample. Other data collection methods, such as internet surveying or personal interviews may also be used, as long as data quality is assured.

Member States may combine the CIS 4 questionnaire with other surveys, **as long as this does not negatively affect the quality of the output of the CIS 4.**

3.4. Data editing

Throughout the processing cycle, there should be a systematic and sustained follow up with the responding enterprises to make sure that the data provided is of good quality and passes all edit checks. Data quality checks have to be done at micro- and macro-level by Member States before the results are finally processed and sent to Eurostat. The checking routines of the SAS programs will be delivered to the Member States.

Of course, the SAS edits can be adapted for other computer systems and Member States can also develop their own checks and edits, i.e. the CIS 4 data could be linked with other national data or be compared with R&D survey data.

4. Data quality

⁶ There are also now procedures available in SAS such as PROC SURVEYSELECT, PROC SURVEYMEANS and PROC SURVEYREG that can perform statistical procedures for complex sample surveys.

4.1. Response rates

The units that do not respond to the CIS 4 survey questionnaire may have different characteristics than those that do respond. Therefore, all efforts shall be made to minimise unit (and item) non-response.

The recommended technique to elicit response is to send at least two reminder letters to the sampled enterprise. These should be sent out within an acceptable period after the sending of the original questionnaire. In some cases, timely telephone reminders may also prove useful.

4.2 Unit non-response and non-response survey

If non-respondents, as an unweighted percentage of all relevant enterprises in the sampling frame, exceed 30%, then a simple random sample of **at least** 10% of the non-respondents (excluding non-relevant enterprises) should be selected. The form to be used for this non-response survey is to be specified. It shall include some of the questions of the standard CIS 4 questionnaire, in order to determine if the non-respondent is an innovator or not. If non-response is not equally distributed across strata, Member States may use a stratified non-response sample.

The non-response survey should have a very high response rate. This non-response survey should be carried out for at least the core target NACE population.

If the results from the non-response analysis indicate that there is a difference between respondents and non-respondents for a certain type of enterprise, this information should be used when calculating the weighting factors (see section 4.5). Member States shall describe how the information from the non-response survey has been used to reduce eventual bias in the estimates.

4.3 Item non response

Item non-response should be kept at a minimum by asking the enterprises for the additional information needed. Item non-response for general variables on the enterprises should not exist, as this information should be available in the business register or from other sources. Some respondents may return questionnaires that have some items filled in, but these cases should only be counted as respondents if they are usable in the processing stage.

Before carrying out automatic imputation, Member States should, as far as possible, make use of administrative, historical (e.g. the CIS 3 survey) or other available data sources such as R&D surveys.

4.4 Imputation

To correct for item non-response (after every attempt is made to get the information from the enterprises concerned) imputations shall be done. Imputed values should be flagged as this enables proper non-response analysis to be done.

The SAS software package (see section 3.1) will impute metric (or measurement) variables separately from ordinal (or ranking) variables, as was done for the CIS 3.

(1) Metric variables

A weighted mean of each metric variable, by NACE and size class, is calculated and applied as a ratio to the enterprises with the missing values, within the stratum concerned.

(2) Ordinal, nominal and percentage variables

This imputation shall be done after the metric estimation. The technique used is nearest-neighbour hot decking using entropy⁷. This technique will use data from clean records (a donor with a record not violating any error check), in order to copy the missing data. The donors are chosen in such a way that the distance between the donor and recipient be minimised⁸.

Member States may also use other reliable methods of imputation, as long as the quality of results is at least identical.

4.5 Weighting and calibration

The survey results should be weighted in order to adjust for the sampling design and for unit non-response to produce valid results for the target population. Additional auxiliary information should also be incorporated, if it is considered that this will enhance the accuracy of the estimates.

The basic method for adjusting for different probabilities of selection used in the sampling process is to use the inverse of the sampling fraction i.e. using the number of enterprises or employees. This would be based on the figure N_h/n_h where N_h is the total number of enterprises/employees in stratum h of the population and n_h is the number of enterprises/employees in the **realised** sample in stratum h of the population, assuming that each unit in the stratum had the same inclusion probability. This will automatically adjust the sample weights of the respondents to compensate for unit non-response.

However, if a non-response analysis is carried out (and the results indicate that there is a difference between respondents and non-respondents), then the results of the non-response analysis should also be used when calculating the final weighting factors. One approach is to divide each stratum into a number of response homogeneity groups with (assumed) equal response probabilities within groups. A second approach could be to use auxiliary information at the estimation stage for reducing the non-response bias.

If the frame contains auxiliary information about the sampling units i.e. variables that are correlated with at least some of the measurement variables of interest, this information should

⁷ Cold deck imputation, on the other hand, makes use of a fixed set of values, which covers all of the data items. These values can be constructed with the use of historical data, subject-matter expertise, etc. A 'perfect' questionnaire is created in order to answer complete or partial imputation requirements.

⁸ Nearest neighbour imputation: In this case a criteria is developed to determine which responding unit is 'most like' the unit with the missing value in accordance with the predetermined characteristics. The closest unit to the missing value is then used as the donor.

be used to improve the estimation further⁹. In general, the variables to use for calibration are turnover and the number of enterprises, both by NACE and size classes but others can also be used.

Various software packages are available to do the calculations needed to derive calibrated weights. These include:

- CLAN. This was developed by Statistics Sweden and it is a suite of SAS-macro commands.
- CALMAR (Calibration on Margins). This is another SAS macro developed by INSEE in France.
- CALJACK. This is also a SAS macro developed by Statistics Canada.

Several different sets of weights may be produced, depending on the variables of interest. In practice however, there will probably be only up to three different weights produced.

Member States are free to use whatever calibration technique they prefer but, in accordance with section 7, paragraph 4 of the annex to the Commission Regulation on innovation statistics, they should provide information about the calibration methods used.

4.6 Precision of results

The CIS 4 should be carried out in order to achieve a certain level of precision for the total population concerning the following indicators:

1. Percentage of innovation active enterprises.
2. Percentage of innovators that introduced new or improved products to the market.
3. New or improved products, as a percentage of total turnover.
4. Percentage of innovation active enterprises involved in innovation cooperation.

These variables are listed in section 1 of the annex of the Commission Regulation on innovation statistics. In addition, the CIS 4 should also achieve a certain level of precision for the total population with regard to the following indicator:

5. Total turnover per employee.

Article 6 of the Commission Regulation on innovation statistics states that quality evaluation shall be carried out by Member States. Therefore, after processing the data,

the 95% confidence intervals¹⁰ for the first three indicators should be $\hat{\theta} \pm 0.05$, for indicator 4 the 95% confidence interval should be $\hat{\theta} \pm 0.10$, and for indicator 5 the confidence interval should be $\pm 10\%$ of the estimate $\hat{\theta}$.

⁹ It can be done for balancing purposes (in the sense that after calibration, “the sample looks like the population”) or for improved consistency of estimates (in production systems, each sampled unit is given a unique final weight as part of the calibration process; as a result, estimates are consistent in the sense that the parts add up to the totals).

¹⁰ The confidence interval for the parameter, $\hat{\theta}$, with approximate confidence level of 95%, is given by:

In accordance with section 7, paragraph 4 of the annex of the Regulation on innovation statistics, Member States shall transmit these quality results to Eurostat.

5. Transmission of data

5.1 Data to be transmitted

Article 5 of the Commission Regulation on innovation statistics lays down two types of data to be transmitted to Eurostat. The first set refers to aggregated statistics that will be transmitted on a compulsory basis while the second refers to individual data records that will be transmitted on a voluntary basis.

The annex to the Regulation says that, beyond the statistics listed in section 1 of the annex, additional tabulated statistics will be decided in close cooperation with Member States. Eurostat will provide the tabulation scheme as well as the transmission format to be used for both data sets (the micro-data set and the tabulated dataset) to Member States.

Aggregated statistics shall be treated in accordance with the standard confidentiality rules at national level (including secondary confidentiality), before transmission to Eurostat. Confidential tabulated data may also however be transmitted, in accordance with Council Regulation 1588/1990¹¹, article 3.

In accordance with section 7, paragraph 4 of the annex of the Commission Regulation on innovation statistics, metadata (which Eurostat will specify) should also be sent. This will include key quality indicators such as non-response rates, coefficient of variation, etc.

The individual data records will be submitted to quality checks. This data will also be used for the compilation of an anonymised micro data set and be made available for further scientific research, according to the procedures laid down in Commission Regulation 831/2002.¹²

5.2 Output tabulation

In accordance with section 5, paragraphs 1 and 2 of the annex of the Commission Regulation on innovation statistics, results will be broken down by economic activity and employment size classes. The output tabulation scheme (which will be produced in accordance with annex 1 of the Commission Regulation on innovation statistics) will be orientated towards the NewCronos CIS3 dissemination structure.

$$\hat{\theta} \pm 1.96 \cdot \sqrt{\text{Variance}(\hat{\theta})}$$

¹¹ Council Regulation 1588/1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities.

¹² Commission Regulation 831/2002 mentions the Community Innovation Survey as one of the surveys where anonymised micro data may be made available to researchers under specific conditions (controlled access).

However, with regard to regional data, the tabulation scheme will also contain results broken down by:

- NUTS 2 level by industry (NACE C to E) and services (NACE G to K).
- NUTS 2 level by size classes (as listed in section 2.3).

5.3 Transmission tools

CIS 4 data shall be transmitted to Eurostat via STADIUM. This safe, secure procedure guarantees a method of tracking transmission. All necessary steps should be taken to ensure that the STADIUM system is working at national level.

5.4 Deadlines

The deadlines for data transmission listed in the annex of the Commission Regulation on innovation statistics should be respected. These deadlines are:

- Transmission of tabulated data – at the latest by 30th June 2006. This will be the main source for data dissemination.
- Transmission of micro data - at the latest by 30th June 2006.

This deadline should also be respected with regard to the transmission of the information related to section 7, paragraph 4 of the annex of the Commission Regulation on innovation statistics i.e. information concerning the methodology used in the national innovation survey.

Annex 1: Target population changes

The following are situations where the target population may change or cause difficulty during the survey:

- Subsidiaries of multinationals requesting contact with the parent organization. While the subsidiaries may get the information from abroad, the information should only relate to the particular national subsidiary. There is a general difficulty with getting multi-national organizations to report information at national level but they will have to make every effort to delineate their data for national units at least. Only domestic units of multi-national corporations should be included in the survey.
- Companies under liquidation or that were liquidated during the observation period (2002-2004 inclusive). Companies that were liquidated before the period should not be considered as part of the target population. Companies that were liquidated during the period should also be deleted from the sample and target population, unless it is decided that their liquidation was so late in the survey period that they should be included in the target population.
- New companies created during the observation period. These should be added to the population.
- Enterprises changing NACE section. These should be recoded accordingly and considered as part of the new NACE section rather than the old one.
- Two or more enterprises combine to form one enterprise. If this happened before or at the beginning of the survey period (and one or more of the units is in the sample) then the new unit should respond with a single form for both (or more) enterprises. Additionally the population should be changed to delete the two (or more) individual units and to include the new unit only. If neither unit was in the sample then the population should simply be amended to reflect the changes.

If the merger happened late in the survey period, then the original units can be treated as they are, i.e. separately, and ignore the merger. Care will have to be taken however that neither unit returns information for more than its' original elements and they do not send in responses covering the other merged elements as well.

- Enterprises that split to form new units. If this happened early in the survey period then the target population should be amended to reflect the new units. Any such enterprise that is part of the sample should return forms for each new unit separately. If the split happens late in the survey period or if the enterprise cannot supply information on each new element separately, keep the unit as it was before the split.
- Enterprises that are outside the target population, i.e. in NACE sections not covered by CIS4. These should be excluded from all processing if they are in the sample. In addition, the target population should be adjusted before the calculation of weights, in order to exclude these and other types of non-relevant enterprises.

Annex 2: Sample size calculation and allocation¹³

Generally, the factors that affect precision of the results are:

- Size of the population
- Variability of characteristics in the population
- Sample plan and estimators
- Non- response
- Cost and time
- Operational constraints (like training of staff etc.)

I. Estimation of parameters

Consider a set of variables $y_1, \dots, y_a, \dots, y_A$ and let $y_a(k)$ be the value of variable y_a for unit k in the finite population U . Also, consider a partitioning of U into D possibly overlapping domains $U_1 \dots U_2 \dots U_D$. For each one of the $A^x D$ possible combinations of variables and domains, a number of parameters θ of interest can be defined for the whole population or for different domains.

II. Sample design

The sample is drawn as stratified sample with simple random sampling without replacement within strata. The stratification is according to section 2.3, taking into account the study-domains for the output tabulation in section 5.2.

III. Sample size in domains of study

Each domain is considered as a population, which is divided into one or more strata. The sample size, n_D , in domain D is calculated as:

$$n_D = \frac{\left(\sum_{h=1}^H W_h \cdot S_h \right)^2}{V(\hat{\theta}_D) + \frac{1}{N_D} \sum_{h=1}^H W_h \cdot S_h^2} \quad (2.1)$$

where $V(\hat{\theta}_D)$ is the variance for the estimated parameter; H is number of strata in domain D ; $W_h = N_h / N_D$, where N_h is the number of enterprises in stratum h ; N_D is the number of enterprises in domain D ; and S_h^2 is the stratum variance for the variable, y_a .

¹³ For general information on sampling, see Cochran W. G. (1977) Sampling Techniques, third edition, John Wiley.

$$S_h^2 = \frac{1}{N_h - 1} \sum_{k \in a_h} \left(y_a(k) - \frac{1}{N_h} \sum_{k \in a_h} y_a(k) \right)^2 \quad (2.2)$$

The expression in (2.1) is obtained by considering the cost to be equal for all strata, e.g. $c_h = c$ for all h , as in formulae (5.25) in section 5.5 in Cochran¹⁴.

IV. Precision

The confidence interval for the parameter, θ , with approximate confidence level of 95%, is given by:

$$\hat{\theta}_D \pm 1.96 \cdot \sqrt{V(\hat{\theta}_D)} \quad (2.3)$$

The precision, α_D , in terms of the length of the confidence interval:

$$\alpha_D = 1.96 \cdot \sqrt{V(\hat{\theta}_D)} \quad (2.4)$$

From (2.4) the variance, $V(\hat{\theta}_D)$, can be expressed as:

$$V(\hat{\theta}_D) = \left(\frac{\alpha_D}{1.96} \right)^2 \quad (2.5)$$

By combining (2.1) and (2.5), the sample size in domain D is given by:

$$n_D = \frac{\left(\sum_{h=1}^H W_h \cdot S_h \right)^2}{\left(\frac{\alpha_D}{1.96} \right)^2 + \frac{1}{N_D} \sum_{h=1}^H W_h \cdot S_h^2} \quad (2.6)$$

Note

1. To calculate n_D , the true variances in each stratum, S_h^2 , is needed and the precision, α_D .
2. In practice, the standard deviations for each stratum, S_h , are not known. Therefore, the CIS 3, CIS Light or other sources might have to be used, but these estimates might be rather unreliable.
3. The above-described sample size calculation will ensure that the sampling error of a specific variable does not exceed the predetermined value. However, in section 4.6

¹⁴ Cochran W. G. (1977), Sampling Techniques, third edition, John Wiley; section 5.5 (Optimum Allocation)

there are 5 indicators for which a certain level of precision should be attained. The sample size thus needs to be calculated for each indicator and the largest sample size should be used.

II. Allocation

If the cost per unit is the same in all strata, then the *Neymann allocation* can be used. The total sample size in the domain, D , is distributed among strata, e.g. the sample size in stratum h , n_h , is given by:

$$n_h = n_D \cdot \frac{N_h \cdot S_h}{\sum_{h=1}^H N_h \cdot S_h}. \quad (2.7)$$

Note

1. The determination of an optimum allocation is often an iterative process. The first step may yield, in some strata, a sample size larger than the number of enterprises in the population. The usual procedure is to take all enterprises in those strata as part of the sample and subsequently reduce the total sample size and recalculate n_h again for the remaining strata.
2. The above-described allocation is optimal for a specific variable. It might not be the case when allocating the sample for other variables and “compromise” allocation schemes are needed. For the CIS4 the sample has to be allocated in order to meet the precision criteria for the 5 indicators for which a certain level of precision of results is required (see section 4.6).
3. Several different such schemes can be used. A simple procedure for multivariate allocation is to compute the average sample sizes for each stratum but methods that are more sophisticated may also be used.

Annex 3: Data editing

The types of checks being done in the SAS programmes are:

- Completeness checks. This is where the questionnaire is not fully completed. Contact should be made with the reporting unit to get the information as soon as possible after receipt of the incomplete form.
- Out of scope units. These are units which do not belong to the target population i.e. wrong NACE, wrong size etc. If this is the case, i.e. if the units are not part of the target population, then they will be dropped from further data processing.
- Data validation checks. This tests whether answers are permissible i.e. the answer is within the range of answers allowed. If a validation error occurs then the answer must be amended (by getting further information from the enterprise for example) to bring it into line with the range allowed.
- Relational checks. This checks that the relationship between two variables is within specific bounds i.e. innovation expenditure should equal the total given. These errors may be “hard” (a violation of the rule indicates that something is incorrect) or “soft” (just a warning that something might be wrong). The hard errors will have to be corrected while the soft errors should be confirmed with the enterprise (and corrected if the information is actually wrong).
- Routing errors. This tests whether all questions that should have been answered have been answered, i.e. innovators answered questions on effects of innovation. An error here indicates that the respondent did not understand the sequencing of questions. They should be contacted to correct the information.

A more complete description of the data editing (and also imputation, estimation etc.) procedure will be provided with the updated SAS programs.

Annex 4: Total Design Method

The Total Design Method (Dillman, D. (1978): *The Total Design Method*, Wiley) consists of a combination of actions (or moments) that have proven effective in reducing non-response when using mail questionnaires.

The theory underlying the TDM is social exchange, which suggests that the likelihood that individuals will respond to a survey questionnaire is a function of how much effort is required to respond, and what they feel they are likely to get in exchange for completing the questionnaire.

The TDM was originally developed for individual and household surveys. An adaptation for the business environment is described in *Tailored Design Method* (Dillman, 2000) and Moore & Baxter (Moore, D. and Baxter, R. 1993) in “Increasing Mail Questionnaire Completion for Business Populations: The Effects of Personalization and a Telephone Follow-up Procedure as Elements of the Total Design Method”.

Five main actions that can be used to improve response rates in business surveys are:

Have a respondent-friendly questionnaire. This should be easy and clear to understand, have a relevant question order and a comprehensible, “user-friendly” layout.

There should be up to five contacts with the potential respondent. A pre-notice letter (sent to respondents a few days prior to the questionnaire), the questionnaire (sent a few days to a week after the pre-notice letter, a thank you/reminder postcard (sent about one week after the questionnaire). If necessary, there should also be a replacement questionnaire (sent to non-respondents between 2-4 weeks after questionnaire was mailed) and a final contact (made a week after the replacement questionnaire was sent out).

In all cases where mail response is requested, the use of a real stamp on return envelopes can increase the response rates (It represents something of value and is something the respondent is less likely to throw away).

Personalised correspondence could be used by using real stationery, real names and real signatures.

Finally, a small token or financial incentive can significantly improve response rates. However, incentives can have modest and, in some cases, no effect at all.

Other references that can be consulted for more information are:

Paxson, M.C.; Dillman, D.A.; Tarnai, J.: *Improving Response to Business Mail surveys*.
Dillman, D.A.: *Mail & Internet Surveys: The Tailored Design Method*. Wiley, 2000

Annex 5: Testing the non-response survey

The aim of this analysis is to sample a selection of non-respondents and find out if they have a different behaviour than that of the original respondents.

If a non-response survey has been carried out (as it should be if the non-response rate is above 30%, i.e. 30% or more of relevant enterprises did not respond to the survey), a statistical test has to be carried out to check whether the population of non-respondents is significantly different from the populations of respondents.

Test for the equality of two proportions:

$H_0: P_R = P_{NR}$ or $P_R - P_{NR} = 0$ where P_R is the weighted percentage of innovators in the respondent population and P_{NR} is the weighted percentage of innovators in the non-respondent population.

$H_1: P_R \neq P_{NR}$

Test statistic:
$$Z = \frac{(\hat{P}_R - \hat{P}_{NR})}{\sqrt{S^2(\hat{P}_R) + S^2(\hat{P}_{NR})}}$$

$S^2(\hat{P}_R)$ is the estimated variance of the proportion of innovators in the original, realised sample, calculated after weighting for sampling fractions while $S^2(\hat{P}_{NR})$ is the estimated variance of the proportion of innovators in the non-response sample.

If a simple random sample or a stratified sample of the non-respondents is drawn then the variance, $S^2(\hat{P}_{NR})$, would be calculated as:

$$S^2(\hat{P}_{NR}) = \sum \left(\frac{N_h(1-r_h)}{N(1-r)} \right)^2 \left(\frac{\hat{P}_{NRh}(1-\hat{P}_{NRh})}{n_{NRh}} \right) \left(1 - \frac{n_{NRh}}{N_h(1-r_h)} \right)$$

Where $\left(\frac{N_h(1-r_h)}{N(1-r)} \right)$ is the weight of stratum h .

\hat{P}_{NRh} is the percentage of innovators in the non-response sample in stratum h

N_h is the total number of units in the frame population in stratum h

n_{NRh} is the number of units in the non-response sample in stratum h

r_h is the response rate of the original sample in stratum h

With large enough sample sizes, the Z-statistics will be approximately normally distributed. Therefore, if the test statistic is in the critical region (usually defined as greater than 1.96 or less than -1.96, for a 95% confidence interval) then H_0 can be rejected i.e. there is a statistically significant difference between the two proportions¹⁵.

¹⁵For further information, see Wonnacott, H., and Wonnacott, J. R., Introductory Statistics, 5th Edition, John Wiley, 1990, chapter 9.

Annex 6: Imputation procedures

The SAS program documentation for CIS 4 describes the process of imputation in more detail. However, a brief description is given here.

Metric imputation

Metric imputation shall take the “clean” data set, estimate the missing items and create a complete metric data set.

The steps involved are:

- Detect and exclude outliers from calculations of the mean.
- Impute the weighted ratio mean, taking into account the amount of missing values within each stratum.

The key factors affecting metric imputation are:

- Values of the three parameters (factor1, factor2 and remout) which control the process
- Amount of item non-response

Factor1 is the outlier value used to remove extreme values from the dataset (of responses for that variable) before imputation. By default, this is 1.5 (or 1.5 times of the inter-quartile range). In a skewed distribution, this might lead to too many records being rejected. This criterion is checked by the value of the Remout variable. By default this is 30, i.e. do not use factor1 where its use leads to the rejection of 30% or more of the records. If the remout value is exceeded, then the imputation procedure moves onto factor2. By default this is set at 3.0 i.e. use all records within 3.0 times of the inter-quartile range.

The three variables controlling the imputation procedure can be amended within the SAS program but, for comparability purposes, it is important that the values used should be as close to the default values as possible. Therefore, the first step to improve item non-response should be to improve response rates. It is very important that item non-response should be kept to a minimum.

After this has been done, if the variables controlling imputation have to be changed (because records are still not being imputed), start off by increasing the remout value little by little until the imputation procedure improves (for example reduce from 30% to 25% to 20%). If this does not work increase factor2 and remout (from its original value) until the imputation procedure produces acceptable results.

If item non-response within a stratum is higher than 50% then the stratum is merged with a neighbouring size class in the same NACE class. If the proportion of non-missing values is still lower than 50% for all size groups within the NACE class the imputation is implemented within subsections of NACE or ultimately by using the whole population. Where strata have non-response rates higher than 50%, every effort should be made to improve the results for these critical strata.

Ordinal and nominal imputation

After the metric estimation comes the Ordinal estimation. The objective of this process step is to estimate nominal and ordinal variables (and in some cases metric variables). As for the metric estimation, it is the amount and structure of the item non-response that is the main factor influencing the outcome of the imputation process.

The basic method is:

- Metric variables are broken down into classes. Respondents are partitioned into classes such that the elements in the same class are considered similar. The variables used here are NACE and size class.
- Metric and ordinal variables are used to estimate nominal variables.

The key factors affecting the ordinal imputation are:

- Values of one parameter (classl) which controls the process
- Amount of item non-response

ClassL determines how much data to include for each variable in the imputation process. If ClassL=2 then only one class is created around the median, excluding large proportions of the data (outliers). ClassL=5 includes more data and creates 4 classes etc.

If there is still item non-response after ordinal estimation, there might be several reasons for this:

- Item response is very low, too low for some strata. This should be addressed by trying to improve response rates in these critical strata at least.
- The setting of ClassL is too strict, reducing the critical mass of data for the estimation procedure. Therefore, increase ClassL to include more data.

However, as for metric estimation, it is important that the final setting is as close to the benchmark (set for each variable in the SAS programs) as possible, in order to maintain comparability of data.