IDEA paper

IDEA PAPER SERIES

IDEA 4 1998

USING *COMMUNITY INNOVATION SURVEY* DATA FOR EMPIRICAL ANALYSIS

- data reliability and issues for analysts

Tore Sandven and Keith Smith

Tore Sandven is a researcher at the STEP group. Keith Smith is research director at the STEP group. STEP group, Storgaten 1, N-0155 Oslo. Tel: +47 22 47 73 10, Fax: +47 22 42 95 33, email: tore.sandven@step.no, keith.smith@step.no.

STEP group

Studies in technology, innovation and economic policy
Studier i teknologi, innovasjon og økonomisk politikk

IDEA

This report is part of Sub-Project 1.1, 'Basic Concepts of Innovation and Problems of Measurement', of the IDEA (Indicators and Data for European Analysis) Project. IDEA is Project No. PL951005 under the Targeted Socio-Economic Research Programme, Area 1 (Evaluation of Science and Technology Policy Options in Europe), Theme 1.3: *Methodologies, Tools and Approaches Relevant for the Preparation, Monitoring and Evaluation of Science and Technology Policies.*

An overview of the project as a whole, covering objectives, work programme, and results, including downloadable reports, can be found on the IDEA Web-site:

http://www.sol.no/step/IDEA/

ABSTRACT

A large part of the IDEA project has been concerned with conceptual and methodological issues related to new indicators for innovation inputs and outputs.

The most important recent initative in innovation indicator development is the *Community Innovation Survey* (CIS). The objective of this report is to identify and discuss interpretive issues arising from sampling problems in the *Community Innovation Survey*. The report focuses on the data on innovation expenditures within the CIS, and shows that there are significant sample biases in this data, and moreover that the biases vary across countries. These problems do not preclude international comparative analysis, but they sharply limit its scope. It will be essential for future rounds of CIS to overcome such problems, particularly by setting and insisting on minimum adequate response rates in each country collecting the data. But it would also be appropriate for a much greater emphasis to be given to all of the myriad issues which are connected with sampling in surveys of this type, especially in future rounds of CIS.

TABLE OF CONTENTS

TABLES

FIGURES

INTRODUCTION

The aim of this report is to assist policymakers and analysts in the use of an important data source, namely the first round of the *Community Innovation Survey* (hereafter CIS): the objective of this report is to identify and discuss indicator problems related to sampling issues in the Community Innovation Survey. This is relevant both to analysis of the data from the first round of CIS, but also to collection methodologies in the second round of CIS, which is taking place in 1997 and 1998.

The *Community Innovation Survey* is a significant project, in at least three ways. Firstly, the data is of a new type. There has never before been such a large-scale attempt to collect internationally comparable data on non-R&D resources devoted to innovation, or data on direct measures of innovation outputs. Secondly, CIS collects and assembles data at firm level, and it makes firm-level data available to analysts. The aim is to give CIS users a view of what is happening at the level of industries (which is the level at which most industrial and R&D data is available), but also to give them a precise and detailed statistical picture of what is happening *inside* European industries. In this sense it is an important source in discussing issues related to variety and diversity within industies. The third innovative feature is the scale and scope of the project. The survey collected data on approximately 200 variables for each firm, and the final CIS database contained records for approximately 40,000 firms. In terms of its scope and coverage, its international dimensions, and the volume of information available, the CIS database is a potentially unique policy resource. For the purposes of this report we will assume that those who read it are broadly familiar with the questionnaire and the main variables.

CIS developed and incorporates data on the following topics:

- expenditure on activities related to the innovation of new products (R&D, training, design, market exploration, equipment acquisition and tooling-up etc). There is therefore a unique focus on non-R&D inputs to the innovation process.

- outputs of incrementally and radically changed products, and sales flowing from these products

- sources of information relevant to innovation

- R&D performance and technological collaboration

- perceptions of obstacles to innovation, and factors promoting innovation

For a full description of these variables, the reader should consult the European Commission document, **The Community Innovation Survey - Status and Perspectives** (Luxembourg 1994).

Scope and focus of this report

In this report we focus on the analysis of innovation expenditures by firms, looking mainly at the potential for meaningful international comparisons across firms and industries. It is well known that R&D indicators do not necessarily give us a good picture of inputs to innovation process across industries, since many firms and industries innovate via such activities as design, engineeering development, and so on. The CIS for the first time collected a large volume of data across countries on this, and it therefore gives us – at least potentially – the possibility to explore inter-industry and inter-country variations in the level and composition of non-R&D inputs to innovation. But the potential for such analysis depends very much on the nature of the data. The bulk of the report is therefore a technical analysis of the characteristics and quality of the data on innovation expenditures. The main problem on which we focus is the existence and implications of various types of sample bias within the data.

The Basic Data

The following table shows the number of firms in each country in the sample and how these firms are divided between those who report innovation activity and those who do not.

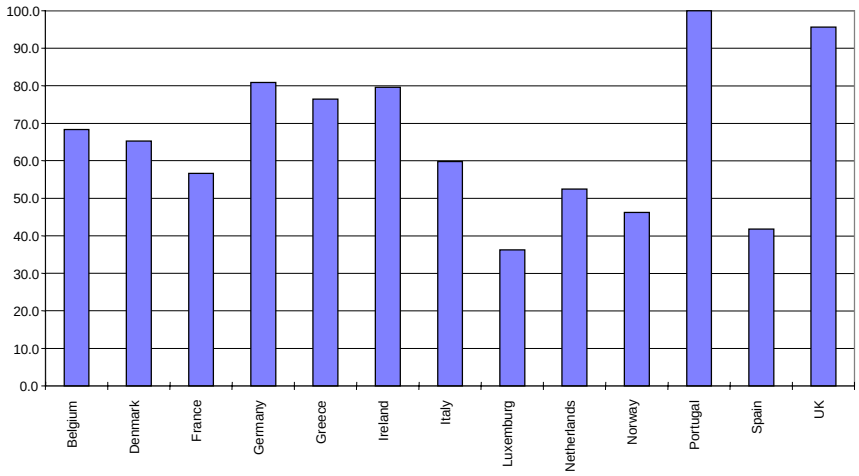
Table 1: Number and share of innovative firms in 13 European countries

Country	*N*	*N innovative*	*N not innovative*	*share innovative*
Belgium	748	511	237	68.3
Denmark	674	440	234	65.3
France	3879	2196	1683	56.6
Germany	2918	2360	558	80.9
Greece	399	305	94	76.4
Ireland	999	795	204	79.6
Italy	22788	13620	9168	59.8
Luxemburg	372	135	237	36.3
Netherlands	4094	2149	1945	52.5
Norway	982	454	528	46.2
Portugal	410	410	0	100.0
Spain	2372	991	1381	41.8
UK	182	174	8	95.6

Let us term those firms who report innovation activity 'innovative', those who do not 'not innovative'. The criterion for being innovative is to have answered 'yes' to at least one of the introductory questions concerning new products and processes (v1, v2 and v3), for being not innovative to having answered 'no' to all three of them. The latter are asked to skip most of the remaining questionnaire, including all questions on innovation costs. Being not innovative thus implies that they have no innovation costs (in 1992, the year under consideration).

In the table we have also shown the share of the firms for each country who are innovative. This is also shown graphically in the figure below.

Figure 1: Share of firms who report innovation activity



We see that there is quite substantial variation across countries in the share of the firms who are innovative.

Problems concerning the comparison of innovation cost intensities

Let us distinguish between two main ways of characterizing innovation costs available to us. One is to look at their level, or more accurately their *intensity*. Here we will choose to express this intensity as the proportion between innovation costs and sales. The other dimension is the *composition* of innovation costs. Let us look at the problem of the *intensities* first.

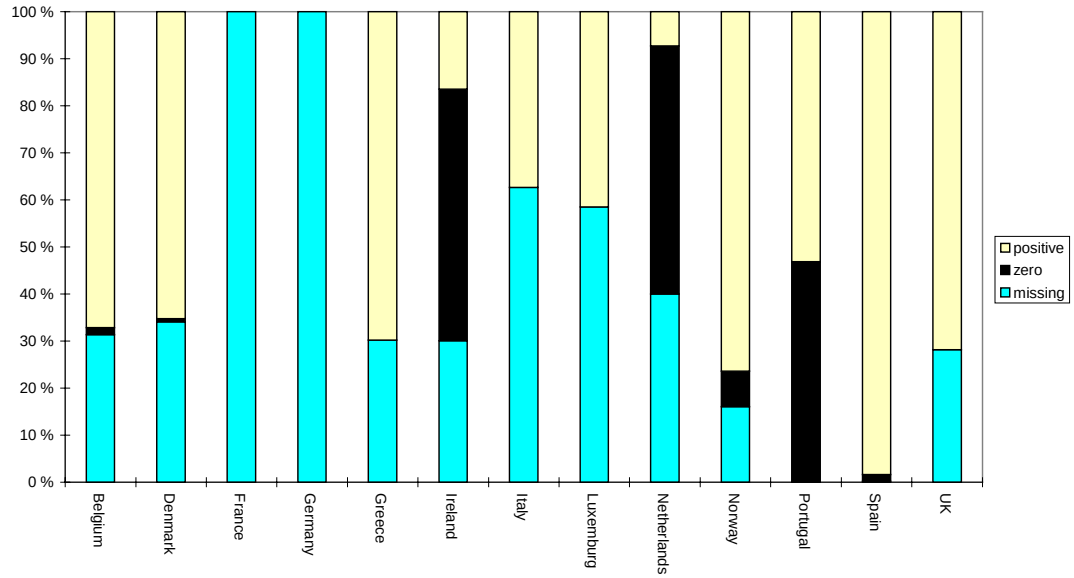
Here we are interested in both central tendency of different distributions, i.e. averages (which may also be *weighted* averages) or other measures such as the median, and in spread, or perhaps especially in the skewness or inequality of different distributions. We are interested in comparing across notably countries, industries and size classes. It is important that the basis for the calculation of averages is comparable across classes. This is also essential for comparing the inequalities of different distributions. One interesting aspect of the variability of innovation cost intensities across classes is the variation in the share of the firms (and the share of total sales which they represent) who have no innovation costs. We shall see that there are large problems establishing comparable bases for comparing across countries here.

We have already seen that the share of the firms who are innovative varies substantially across countries. Thus, also the share who are not innovative varies across countries. We know that the latter firms have no innovation costs.

We now have to turn to the firms who are innovative. Some of these report some positive value for innovation costs, some explicitly report the value zero, and for some firms the value is missing. In the following three figures are shown for each country: the share of the innovative firms who have, respectively, reported no answer, explicitly reported zero and reported a positive value for innovation costs. There is one figure for each of three types of innovation costs: current innovation costs, investment costs related to innovation, and R&D costs.

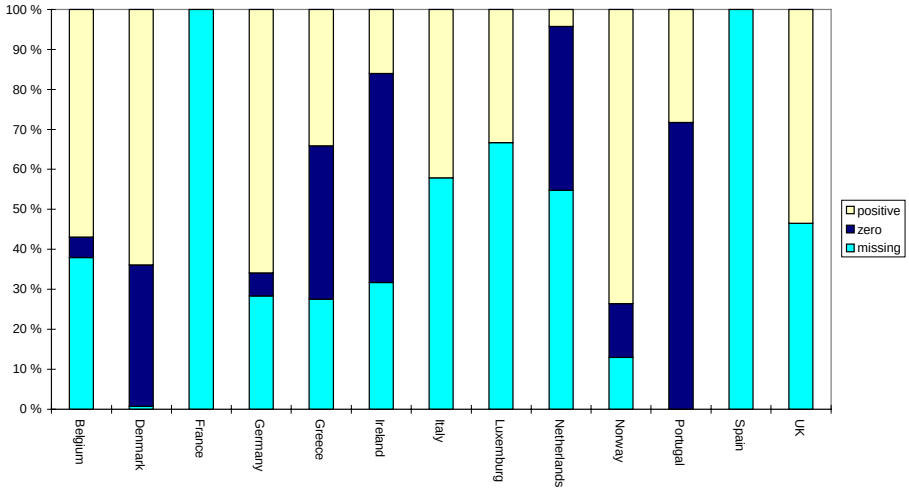
The first of the three figures shows the distribution for current innovation costs (v13a).

Figure 2: Current innovation costs: distribution of innovative firms



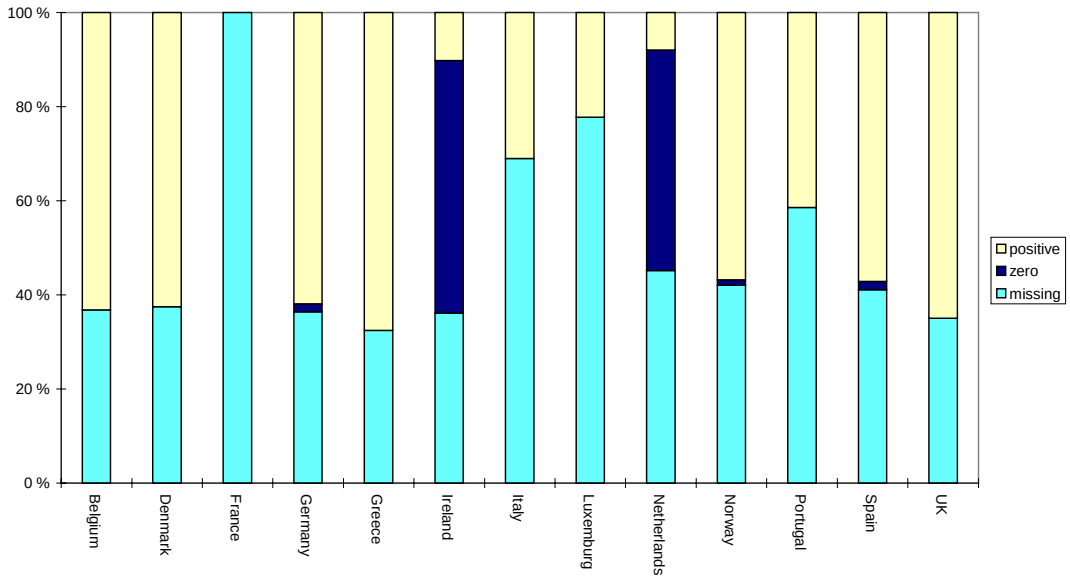
Next we turn to the distribution for investment costs related to product innovation (v13d).

Figure 3: Innovation investment costs: distribution of innovative firms



Lastly, we turn to R&D costs. R&D costs are a component of the current innovation costs (from v13a) reported above, but there is also a special group of questions in the questionnaire concerning R&D activity, and it is the data on R&D costs (v10d_1) on which the following figure is based.

Figure 4: R&D costs: distribution of innovative firms



These figures show, in our view, that missing values are a very serious problem here. For one thing, France falls out of the picture altogether, while for Germany there are no data on current innovation costs and for Spain there are no data on investment innovation costs.

Apart from this, the share of missing values is very substantial, in most cases accounting for at least 30 per cent of the innovative firms. But more serious than this is the very large *variability* in this share across countries, making comparison across countries very difficult.

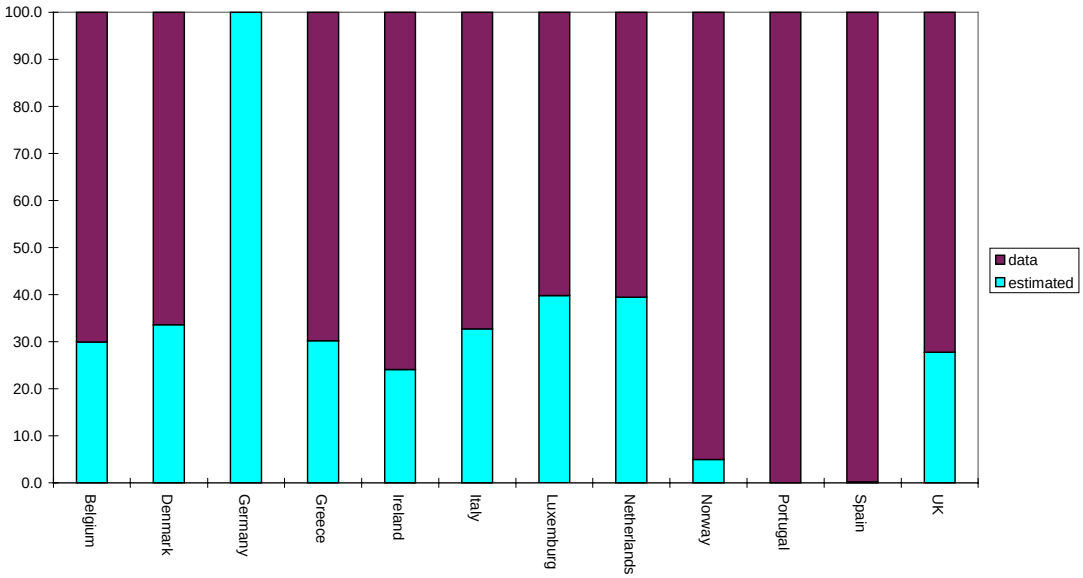
A special, and very serious, problem is the variation across countries in the extent to which there is distinguished between the value zero and a missing value in the data. For some countries, notably Italy, Luxemburg and the UK, and for most countries in the case of R&D costs, there are no zero values, all cases where there is not reported a positive value evidently having been registered as missing. On the other hand, in other cases (Portugal and Spain for current costs, Denmark and Portugal for investment costs), all cases where there is not reported a positive value have evidently been registered as zero. In many cases, both zero values and missing values are registered, but with very large variation in the share of zero values across countries, both in relation to all innovative firms and in relation to all non-missing values. These differences across countries seem far too large to be credible.

A word may be said about the case of zero innovation costs for firms who report innovation activity, which means that they report either product or process innovation in the course of 1990-92 or the intention of introducing product or process innovations during 1993-95. It might be claimed that innovation costs here are defined so widely that it is not possible to make changes in products or processes without incurring some kind of additional expenses which should have been reported as innovation costs. If this should be the case, the registering of zero *current* innovation costs is almost an inconsistency. we say almost, because an innovation introduced in 1990 or 1991, although requiring expenses in these years or earlier, does not require any innovation expenditures in 1992, which is the year which applies for the expenditures in the questionnaire. On the other hand, one might claim that innovation costs are not so widely defined, and that it is possible to make significant changes in products or processes as a by-product of the normal activity of the firm, through learning by doing, etc. In any case, a claim concerning the impossibility of zero innovation costs for an innovative firm can at best be valid for *current* innovation costs. Clearly, one can have innovations in the sense defined without having *R&D* expenditures, and a certain share of the innovative firms in each country should definitely have the value zero here.

In any case, also, the share of the innovative firms in the Netherlands who report zero innovation costs, all cost categories, seems far too high to be credible. This especially applies if we relate the number of zero values to all non-missing values.

We now want to comment on the estimations that Eurostat have made on the missing values. This is in effect a digression, because we think these estimates are totally unusable. In the figure below are shown the share of data and of estimated values for current innovation costs after the modifications made by Eurostat.

Figure 5: Current innovation costs, non-missing values, after estimation by EUROSTAT: distribution between data and estimated values



We see that the data for current innovation costs, after modification by Eurostat, contain a very high share of estimated values. In the case of Germany, the share is 100 per cent, 2349 estimates having been made on the basis apparently of no data on current innovation costs. But also if we disregard Germany, the share of estimated values is generally very high. We should note that not all missing values have been estimated, far from it.

The estimations have been made by linear regression, or, where the R^2 was considered too low, by registering the mean of the size class and NACE code. However, we know that there is very large variation in innovation cost intensities also *within* industries and size classes and it is generally difficult to get high R^2s between innovation cost intensities and other variables. This state of affairs is somewhat obscured by the example which Eurostat itself uses to explain the

estimation procedure, where sales in 1990 in large enterprises in the Netherlands is estimated from sales in 1992 and employees in 1992 and some third variable, and where the R^2 is as high as 0.97.

Concerning the estimation, it is also difficult to see that any consistent criterion has been applied as to how large a share of the missing values have been estimated. This seems to vary rather arbitrarily across countries.

What would also have been a problem if we were to use the estimated data is that nothing much has been done with the problem of distinguishing between zero and missing. By and large, Eurostat just reproduces the state of the data here. Where the data do distinguish a zero category from the missing category, some of the missing values have been estimated to zero and some to positive values. Where the data do not distinguish between zero and missing, all estimations have positive values. But this rule is not consistently applied either, it appears. Notably, in many cases where there are zeros in the data, all estimated values are nevertheless positive. This especially applies in the case of innovative investment costs.

But in any case this is a digression because no matter how competently the estimation had been done, it may not have been appropriate to use the estimates. It is not only a question of filling in a few holes in the data, since there are far too many missing values. Bit also, the variation in the innovation costs inside classes is too large.

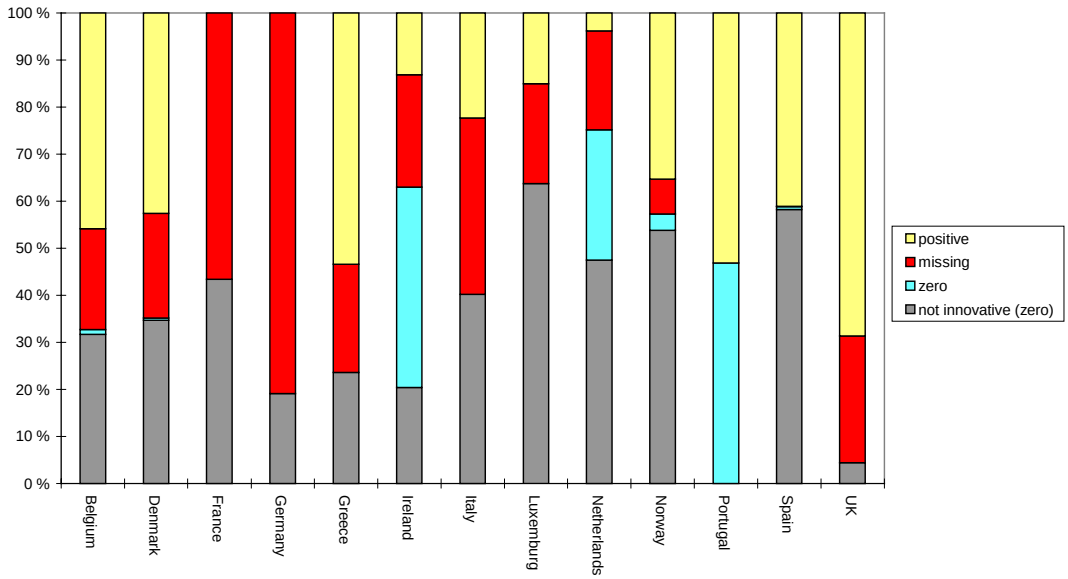
To conclude this digression, it is best for analysts to use only the original data, deleting the Eurostat estimates (apart from where the estimates are intended as simple logical corrections, to be discussed further below). Estimation can be no solution to the missing data problem.

Thus, missing data pose a serious problem to the analysis of innovation costs. We see three aspects of this problem. First, the share of missing values is generally very high. Second, and perhaps even more seriously, there is substantial variation in this share across countries. Third, and equally seriously, there is very substantial

variation across countries in the extent to which the data distinguish between the value zero and a missing value.

Let us now look at the picture which emerges when we consider *all* firms, not just the innovative ones. we will here focus on current innovation costs only. The figure below shows the distribution of all firms by country on the categories not innovative (= no innovation costs), zero, missing and positive.

Figure 6: Current innovation costs: distribution of all firms



There are apparently two unproblematic groups of firms here. One is the one comprising those who report no innovation activity, and who therefore have no innovation costs. The other, obviously, is the one comprising those who have reported a positive value. Then there is a large, highly variable across countries and thus very problematic category of those who have missing values. Then there is the zero category which is also very problematic because it is obviously not applied in the same way across countries, in several countries not even being distinguished from the zero category. Besides, at least in the case of the Netherlands we have good reason to suspect the share of firms who report zero costs as far too high.

But let us go one step further still. We now want to raise serious doubts also on the comparability across countries of the figures for the shares of the firms who are innovative. This means that even if the problem of missing values was solved, there would remain serious difficulties connected to comparing, for instance, the share of firms with innovation costs across countries, the weighted average innovation cost intensities across countries, different measures of inequality of distributions across countries, etc.

As we saw in the beginning of this note, the share of firms who are innovative (who have answered 'yes' to at least one of the introductory questions) varies considerably across countries. Does this reflect real differences across countries, or is it partly an artefact of the implementation of the survey itself?

One striking aspect of the implementation of the survey is the very large variation in *response rates* across countries (documented in 'Evaluation of the Community Innovation Survey (CIS) - Phase I' by D. Archibugi et al., see for instance p. 87). There is defined a 'gross sample' of firms who have received the questionnaire. When we divide the number of firms in our sample with this gross sample, we get a rough measure of the response rate. The numbers are shown in the table below.

Table 2: National response rates

Country	*gross sample*	*realized sample (N)*	*response rate*
Belgium	1949	748	38.4
Denmark	1313	674	51.3
France	5245	3879	74.0
Germany	13320	2918	21.9
Greece	1799	399	22.2
Ireland	3032	999	32.9
Italy	35182	22788	64.8
Luxemburg	470	372	79.1
Netherlands	8221	4094	49.8
Norway	1882	982	52.2
Portugal	1767	410	23.2
Spain	18002	2372	13.2
UK	4998	182	3.6

Now, consider the following set of very simple hypotheses concerning the response of firms to this kind of survey. On the one hand, we may imagine general variables,

with different values across countries, influencing the rate of response in each country. This will partly be cultural and ideological variables, having to do with the commitment of firms to norms which say that one shall answer to such surveys, etc. Partly they will reflect differences in the implementation of the survey across countries: some agencies may simply be doing a better job than others, some may have more resources and competence than others, some may have sanctioning powers which the others lack, etc.

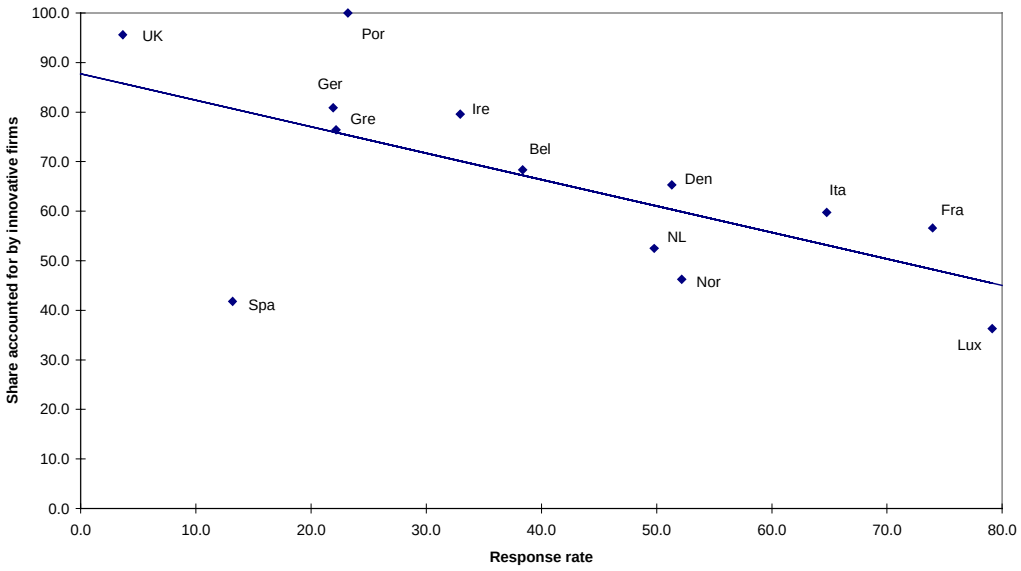
On the other hand, all other things held constant, there will be a tendency for the response rate to increase with the interest which the responsible persons in the firm have in the questions in the survey. Where the questions say them nothing, they will be inclined to throw the questionnaire away, where they find them interesting, they will be more inclined to answer. Now, we also assume a tendency where it is much more likely that people working in a firm which engages in innovative activity will find these questions interesting than persons working in a firm which does not engage in innovation activities. This hypothesis means that the firms who have responded to the survey are likely to differ in a crucial respect from the non-respondents, namely by being more likely to be innovators than the non-respondents.

However, the extent to which this biasing effect is allowed to be operative will precisely be dependent upon the first set of factors influencing the general response rate. Where these factors are so strong that the response rate is 100 per cent, the biasing effect will obviously not be allowed to be operative at all. Where the response rate is high, say 80 per cent, the bias will only be mild. However, where the general factors are very weak and the response rate low, the bias will be substantial.

These assumptions translate into a very simple hypothesis: there will be a clear *negative* relationship between response rate and the share of the firms accounted for by innovative firms. Let us test this hypothesis.

The relationship between these two variables is shown in the figure below.

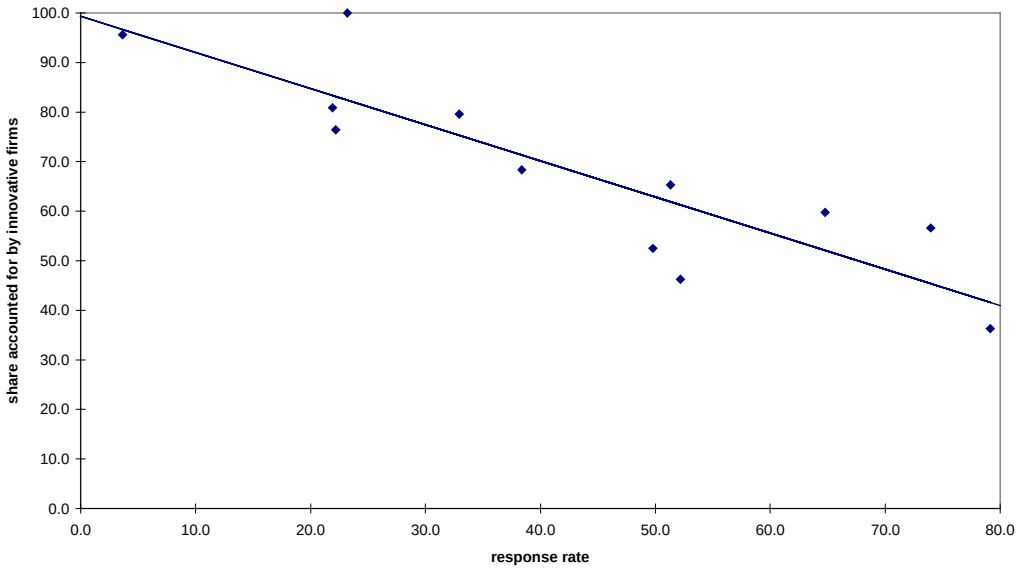
Figure 7: Response rate (x-axis), and share accounted for by innovative firms (y-axis)



As we see, there indeed *is* a negative relationship between response rate and share of innovative firms. The regression line shows this clearly. The correlation between the two variables is given by r = - 0.64, which is significantly different from zero at the 5 % level using a one-tailed test. The lower boundary of *r* of a one-tailed 95 % confidence interval is - 0.23. The regression estimate of the share of innovative firms for the UK response rate, the lowest in the sample, is 85.8 per cent, while for the Luxemburg response rate, the highest in the sample, it is 45.5 per cent.

Spain is clearly an outlier in the above figure. The figure below gives an impression of what the relationship would have looked like if Spain was excluded.

Figure 8: Response rate (x-axis), and share accounted for by innovative firms (y-axis), Spain excluded



Excluding Spain, we get r = - 0.88, which is a very strong relationship.

However, even including Spain we get a fairly strong negative relationship between response rate and the share of innovative firms. And we should bear in mind that this relationship is a bivariate one only. It is possible that we could approach something akin to an adjustment of the share of innovative firms for variables having to do with differences in sample bias if further variables were included. An obvious thing would be to take into consideration the representativeness of the gross sample in the first place. For instance, while in most countries some sort of scientific sampling procedure has been implemented, in Greece and Portugal there is a sample of 'likely innovators' (in the Portuguese case apparently of extremely likely innovators, as the share of innovators is reported to be 100 per cent).

We shall not go further into this here. What is important is that we think the relationship depicted in the above figures constitute fairly convincing evidence that the samples not only are biased in a very crucial respect, but even more importantly, that the degree of bias varies across countries.

Now, what implications should this have for the analysis of innovation expenditures using CIS? We think that the variation across countries in the bias of the sample, together with the high share of missing values and the variation in this share across countries, problems which are confounded by the varying degrees across countries to which the data distinguish between zero and missing values, imply that we cannot in any meaningful way compare the share of the firms who have innovation costs across countries. To us it seems that the best thing which might be done here is simply to focus only on the firms who report innovation costs and calculate all measures of central tendency and dispersion or inequality on the basis of these firms only.

If we choose to do this, we will obviously be restricting our analysis to an *elite* of firms in each country, that is an elite of firms from the point of view of innovative activity. The problem with this, however, is that this elite will not be comparable across countries either. On the contrary, it will represent a more or less restricted elite depending on the country, the sample of firms who report innovation costs perhaps being representative of the top third (relative to innovation activity performance) of firms in one country but only of the top tenth of the firms in another. And what is more, we have, as far as we can see, no way of finding out about this.

Given this state of affairs, one might find that comparing central tendency *across countries,* e.g. comparing averages in each industry across countries, will be of limited interest. The same might be said for measures of spread or inequality. In other words, in relation to central tendency and dispersion, *country* will be very problematic to use as an independent variable. However, it should be far less problematic to use country as a *control* variable, i.e. we can compare across industries and size classes, controlling for country. Thus, given the state of the data, we might perhaps focus more on comparison across industries and size classes and far less on comparison across countries than we originally intended.

Analysis of composition of innovation costs

Let us now turn to the *composition* of innovation costs. We might perhaps find the limitations of the data less serious in the case of the analysis of the composition of the costs than in the case of the analysis of the intensity of the costs. After all, when analyzing the composition of the costs, we have to restrict ourselves to the firms who

have innovation costs. Nevertheless, the difference in the representativeness of the firms across countries remains a big problem.

There are different ways of examining the composition ofinnovation expenditures. For one thing, we have data on both the current innovation costs and the investment costs linked to innovation, or investment innovation costs. Thus, we can look at the composition of total innovation costs between these two categories only.

Secondly, the firms are asked to report how *current* innovation costs are divided among six sub categories. Thus, we can look at the composition of *current* innovation costs across these six categories.

Third, and most comprehensively, we can look at the composition of *total* innovation costs across seven categories, namely, the six sub categories of the current innovation costs, plus investment innovation costs.

Fourth, one interesting division may be to divide *total* innovation costs into *three* categories, namely (1) R&D (current) innovation costs, (2) current innovation costs other than R&D (let us call these non R&D costs) and (3) investment innovation costs.

But let us now look at what the data will allow. W will first deal with the composition of *current* innovation costs. This information is contained in six variables, from v13b_1 to v13b_6. The firms are asked to estimate the share of total *current* innovation costs attributable to each of these six sub categories. They are: (1) R&D, (2) acquisition of products and licences, (3) product design, (4) trial production, training and tooling up, (5) market analysis (excluding launch costs) and (6) other.

There are firms who have answered the question on the amount of current innovation costs by reporting a positive number but who have not given any answer to how these costs are distributed across the six sub categories. Thus the number of firms for which we have data on the composition of *current* innovation costs is lower than the number for which we have (positive figures) data on the amount of current

innovation costs, but not by much, as not very many firms have given a positive figure for the amount without estimating the distribution of the costs.

As above, we think we should define as missing the data where the figures originally were missing but where Eurostat has estimated the composition. On the other hand we think we should accept the estimation where it is only meant to be a logical correction of the data. This is where the shares for the six sub categories does not sum to 100 and Eurostat has simply corrected this.

We should notice that there may be a problem here too, though. The operation seems very simple: where the six reported shares do not sum to 100 but to some figure X different from 100, one should simply multiply ach of the six shares by 100/X to get the logically corrected shares. However, according to the documentation supplied by Eurostat on their own estimations (variables q13b_1 - q13b_6), the logical data correction has invariably been performed on the R&D category (v13b_1) only, whereas nothing apparently has been done to the other component shares. This should mean that the five other shares are accepted as they stand and that the share for the R&D component is estimated by subtracting the sum of the five other shares from 100. Hopefully, this is no more (also: no less) than a case of sloppy documentation. If, on the other hand, this documentation is an accurate description of what has been done, it seems very strange indeed.

Let us now briefly look at the data for the composition of current innovation costs country by country. The following tables, one for each country, are simply the output of *proc means* in SAS, giving N, mean, standard deviation, minimum score and maximum score for each of the six component variables. Notice that the means are normal, unweighted averages.

We start with Belgium.

BELGIUM

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
V13B_1	335	0.4840000	0.3176327	0	1.0000000
V13B_2	335	0.0165970	0.0656547	0	0.7100000
V13B_3	335	0.1145672	0.1824247	0	1.0000000
V13B_4	335	0.2210448	0.2266096	0	1.0000000
V13B_5	335	0.0623284	0.1050945	0	0.8600000
V13B_6	335	0.1012537	0.2079374	0	1.0000000

We see that for Belgium there are 335 firms for which we have data on the composition of current innovation costs. R&D has the highest average share with 48.4 per cent.

We next turn to Denmark.

DENMARK

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
V13B_1	274	0.4021898	0.3158597	0	1.0000000
V13B_2	274	0.0554380	0.1369580	0	1.0000000
V13B_3	274	0.1653650	0.2039789	0	1.0000000
V13B_4	274	0.2620803	0.2422552	0	1.0000000
V13B_5	274	0.0798905	0.1239268	0	1.0000000
V13B_6	274	0.0347445	0.1228542	0	1.0000000

For Denmark there are 274 firms for which there are data. Again, R&D has the highest average share, this time with 40.2 per cent. The next country is Greece.

GREECE

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	207	0.5290338	0.3588223	0	1.0000000
V13B_2	207	0.0651208	0.1744391	0	1.0000000
V13B_3	0	.	.	.	
V13B_4	207	0.2912077	0.3092694	0	1.0000000
V13B_5	207	0.1142512	0.1884332	0	1.0000000
V13B_6	0	.	.	.	

Here there is trouble. We have data for 207 firms. However, apparently neither the product design category (no. 3) nor the residual 'other' category (no. 6) have been applied, as all firms have missing values here. For all firms (i.e. also for each single one) the sum of the remaining four categories is 100. Of these remaining categories R&D has the highest average share, with 52.9 per cent.

We next turn to Ireland.

IRELAND

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	116	0.2462069	0.3135981	0	1.0000000
V13B_2	116	0.0486207	0.1238926	0	0.7700000
V13B_3	116	0.2370690	0.2553736	0	1.0000000
V13B_4	116	0.3578448	0.3168755	0	1.0000000
V13B_5	116	0.1111207	0.1472813	0	0.8400000
V13B_6	0	.	.	.	.

Again there is trouble. There are 116 firms for which there are data. However, the residual category (no. 6) has apparently not been applied, as all firms have missing values here. The remaining shares sum to 100. Of these remaining categories, trial production, training and tooling up has the highest average share, with 35.8 per cent.

Next is Italy.

ITALY

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	5082	0.4216844	0.3296619	0	1.0000000
V13B_2	5082	0.0382054	0.1415075	0	1.0000000
V13B_3	5082	0.2486049	0.2758380	0	1.0000000
V13B_4	5082	0.2357674	0.2562923	0	1.0000000
V13B_5	5082	0.0555864	0.1145783	0	1.0000000
V13B_6	0	.	.	.	.

Also here we have problems, and this is very unfortunate since Italy has by far the largest sample of firms. There are 5082 firms for which there are data, but the residual category ‘other’ has not been applied in this case either. The shares of the remaining categories sum to 100. Of these, R&D has the highest average share with 42.2 per cent.

We now turn to Luxemburg.

LUXEMBURG

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	46	0.2947826	0.3797411	0	1.0000000
V13B_2	46	0.0608696	0.2117632	0	1.0000000
V13B_3	46	0.0978261	0.1594562	0	0.6500000
V13B_4	46	0.2995652	0.3377307	0	1.0000000
V13B_5	46	0.0454348	0.1279706	0	0.7200000
V13B_6	46	0.2013043	0.3511354	0	1.0000000

Here there are apparently no specific problems, but there are only 46 firms. Trial production, training and tooling-up has the highest average share with 30.0 per cent.

The Netherlands is next.

NETHERLANDS

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	125	0.5013600	0.3516192	0	1.0000000
V13B_2	125	0.0345600	0.0937869	0	0.8800000
V13B_3	125	0.0690400	0.1306153	0	0.7600000
V13B_4	125	0.1948000	0.2100676	0	1.0000000
V13B_5	125	0.0767200	0.1167212	0	0.8500000
V13B_6	125	0.1232800	0.2397000	0	1.0000000

There are 125 firms for which there are data. R&D has the highest average with 50.1 per cent.

Norway is next.

NORWAY

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	345	0.3381159	0.3386047	0	1.0000000
V13B_2	345	0.0428986	0.1000770	0	1.0000000
V13B_3	345	0.1430725	0.1941390	0	1.0000000
V13B_4	345	0.3109275	0.3109677	0	1.0000000
V13B_5	345	0.0524348	0.1597905	0	1.0000000
V13B_6	345	0.1122319	0.2206385	0	1.0000000

For Norway we have data for 345 firms. R&D has the highest average share with 33.8 per cent.

Next we turn to Portugal.

PORTUGAL

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	190	0.2415789	0.3450797	0	1.0000000
V13B_2	190	0.0380526	0.1297939	0	1.0000000
V13B_3	190	0.2484737	0.2955978	0	1.0000000
V13B_4	190	0.2694211	0.2859389	0	1.0000000
V13B_5	190	0.0498947	0.1103050	0	0.8600000
V13B_6	190	0.1523684	0.2998115	0	1.0000000

For Portugal we have data for 190 firms. Trial production, training and tooling up has the highest average share with 26.9 per cent.

The next country is Spain.

SPAIN

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	975	0.3643590	0.3877312	0	1.0000000
V13B_2	975	0.0803282	0.2029932	0	1.0000000
V13B_3	975	0	0	0	0
V13B_4	975	0.2189641	0.4132786	0	1.0000000
V13B_5	975	0.0887179	0.1847738	0	1.0000000
V13B_6	975	0.2476103	0.3169797	0	1.0000000

Here there is serious trouble again. We have data for 975 firms, but the product design category (no. 3) has apparently not been applied since no firm has registered any product design costs (although the value zero has been entered instead of missing). The shares of the remaining categories sum to 100, with R&D having the highest average share with 36.4 per cent.

As we have seen, there are additional problems with Spain when it comes to the composition of innovation costs as there are no data on innovation *investment* costs. This means that we have to include Spain when we look at the composition of *total* innovation costs.

Lastly, there is the UK.

UK

Variable	N	Mean	Std Dev	Minimum	Maximum
V13B_1	122	0.3346721	0.3133271	0	1.0000000
V13B_2	122	0.0199180	0.0597307	0	0.4100000
V13B_3	122	0.2700820	0.2838650	0	1.0000000
V13B_4	122	0.2221311	0.2509230	0	1.0000000
V13B_5	122	0.0912295	0.1804429	0	1.0000000
V13B_6	122	0.0618033	0.2077905	0	1.0000000

Although the UK is a highly problematic case owing to the extremely low response rate, there are no specific problems with these data. There are data for 122 firms. R&D has the highest average share with 33.5 per cent.

Thus there are specific problems concerning the data on the composition of current innovation costs for four countries: Greece, Ireland, Italy and Spain. In the cases of Ireland and Italy the components of current innovation costs do not include the residual category ‘other’, while in the case of Spain they do not include ‘product design’. In the case of Greece *both* these categories are missing. For the other countries the average share of the ‘product design’ category varies from 6.9 per cent to 27.0 per cent, while the average share of the ‘other’ category varies from 3.5 per cent to 20.1 per cent.

Composition including also investment costs

We now turn to the composition of *total* innovation costs across all six current cost categories plus the innovation investment category. To have data here, we should require that there is registered a positive amount (a figure higher than zero) for both current innovation costs and investment innovation costs *and* that there are data for the composition of current innovation costs. It turns out that these requirements together reduce the number of firms in the sample considerably. The following table shows, for each country, the total number of firms in the sample, the number of innovative firms, the number of firms for which there are data on the composition of current innovation costs and the number of firms for which we can give the complete composition of total innovation costs, including investments related to innovation.

Table 3: Total number of firms, number of innovative firms, number of firms for which there are data on the composition of current innovation costs and the number of firms for which we can give the complete composition of total innovation costs, including investments related to innovation

Country	*(1) N*	*(2) N innovative*	*(3) N data on composition current innovation costs*	*(4) N data on composition total innovation costs*	*(5) Share data on composition total innovation costs out of total N (%)*
Belgium	748	511	335	250	33.4
Denmark	674	440	274	232	34.4
France	3879	2196	0	0	0
Germany	2918	2360	0	0	0
Greece	399	305	207	99	24.8
Ireland	999	795	116	60	6.0
Italy	22788	13620	5082	3841	16.9
Luxemburg	372	135	46	32	8.6
Netherlands	4094	2149	125	37	0.9
Norway	982	454	345	281	28.6
Portugal	410	410	190	91	22.2
Spain	2372	991	975	0	0
UK	182	174	122	78	42.9

The figures in column 4 show that the number of firms which we end up with here is quite disappointing. Notice also the extremely variable share across countries which these firms account for out of total number of firms in the sample in each country, reported in column 5. Again, this share is the result of a confusing mixture of influences. In addition to the unknown substantive differences these include the variation across countries in the bias of the sample and the varying shares of missing values and zero values.

It should be noted that the number of firms in column 4 has also been limited by the condition that there should be a non-missing value for the NACE code, but the number of firms excluded because of this condition is totally marginal.

Now, of course the requirement that both current and investment innovation expenditures should be reported higher than zero is a strict condition. Could we not relax the conditions somewhat here and accept zero as a response on either current or investment innovation costs if there is reported a positive value on the other category? However, there will be obvious problems with comparability across countries if we do this. We will then have, for instance, that a certain share of the firms in some countries have only current innovation costs while the share of investment innovation costs is zero, whereas this is simply not possible in other

countries, for instance Italy, the firms which actually have there innovation costs composed in this way having already been excluded because of our inability to distinguish zero from missing here.

CONCLUSIONS

We would summarise the conclusions of this analysis as follows:

➨ The samples within CIS are probably biased, in the sense that the sample for most countries probably contains a higher share of innovative firms than the population.

➨ More importantly, the extent of this bias probably varies considerably across countries.

➨ There is a very high share of missing values on innovation costs.

➨ This share varies considerably across countries.

➨ The degree to which the data distinguish between zero values and missing values varies across countries.

➨ All this makes it extremely difficult to compare the share of the firms accounted for by firms with innovation costs across countries.

➨ In particular, we will have to downplay considerably comparisons across countries. Instead we should focus more on comparisons across industries and size classes, using country as a control variable.

➨ The above problems might be thought to be not as serious for the analysis of the *composition* of innovation costs as they are for the analysis of their levels or *intensities*. Nevertheless, important problems remain also here.

➨ The very high share of missing values, especially the shares which result when we combine variables, makes the sample quite small for many countries. This means that we get problems with the proportion of cases to categories.

➨ Our general concusion from this examination of the data is that there are undoubtedly serious problems in the CIS dataset as a result of differences in sampling technique and response rates. These problems do not preclude international comparative analysis, but they sharply limit its scope. It will be essential for future rounds of CIS to overcome such problems, particularly by setting and insisting on minimum adequate response rates in each country collecting the data. But it would also be appropriate for a much greater emphasis to be given to all of the myriad issues which are connected with sampling in surveys of this type.