ECONOMICS – WORKING PAPERS 2020/08

EIB Group Survey on Investment and Investment Finance

A technical note on data quality

European Investment Bank
The EU bank

EIB Group Survey on Investment and Investment Finance

A technical note on data quality

EIB Group Survey on Investment and Investment Finance: A technical note on data quality

EIB Working Paper 2020/08
September 2020

Authors
Philipp-Bastian Brutscher (European Investment Bank)
Andrea Coali (Bocconi University)
Julie Delanote (European Investment Bank)
Peter Harasztosi (European Investment Bank)

This is a publication of the EIB Economics Department

About the EIB Economics Department

The mission of the EIB Economics Department is to provide economic analyses and studies to support the Bank in its operations and in the definition of its positioning, strategy and policy. The department, a team of 40 economists, is headed by Director Debora Revoltella.

economics@eib.org
www.eib.org/economics

Acknowledgements
We thank Pedro de Lima and Christoph Weiss for detailed comments and suggestions.

Disclaimer

The views expressed in this publication are those of the authors and do not necessarily reflect the position of the European Investment Bank.
EIB working papers are designed to facilitate the timely exchange of research findings.
They are not subject to standard EIB copyediting or proofreading.

pdf: QH-BK-20-008-EN-N ISBN 978-92-861-4718-0 ISSN 2599-736X DOI 10.2867/772584

EIB Group Survey on Investment and Investment Finance

– A technical note on data quality

Philipp-Bastian Brutscher (European Investment Bank)

Andrea Coali (Bocconi University)

Julie Delanote (European Investment Bank)

Peter Harasztosi (European Investment Bank)

Abstract

This paper reviews the data quality of the EIB Group Survey on Investment and Investment Finance (EIBIS). It finds that the chosen sampling framework (the Bureau van Dijk ORBIS database) captures the business population of interest well and that there is little evidence of selection bias during fieldwork, suggesting that EIBIS is a reliable data source to study the corporate investment situation in the EU. This result is predicated on the following observations: 1) the ORBIS database has sufficient coverage relative to the actual population; 2) a benchmarking exercise of the final survey sample against randomly drawn samples from the sampling frame shows there is no systematic sampling bias in EIBIS. Efforts to create firm panel do not jeopardize randomness. 3) A comparison of the final sample with two other databases: the Eurostat Structural Business Statistics as well as the CompNet database shows that EIBIS portrays both cross-country differences and dynamics of key variables in a satisfactory way.

Introduction

This paper reviews the data quality of the EIB Group Survey on Investment and Investment Finance (EIBIS), a survey of investment considerations and decisions by European and US firms undertaken annually by the European Investment Banks since 2016.

As with any other survey, the quality of EIBIS depends on two factors: i) the sampling frame from which survey respondents are drawn; and ii) the diligence with which fieldwork is carried out and so non-response and other types of selection bias avoided.

If survey respondents are drawn from a sampling frame that comprises only a sub-sample of the population of interest, e.g. in our case only firms of a particular size or sector, it will be impossible for the survey results to give a representative picture of the investment situation overall. Similarly, if the selected entities are not randomly drawn from the sampling frame or if fieldwork is carried out in a way that allows interviewers to content themselves to speak only to firms that are easily reachable, the final sample of firms will quite likely differ from the underlying population profile. This, in turn, will affect the ability to draw conclusions from the survey sample as representing the broader business population in question.

This paper assesses both of these aspects. After a short presentation of the survey, we examine how well the EIBIS sampling frame, the Bureau van Dijk (BvD) ORBIS dataset, captures the population of interest. In a second step, we assess the quality of the final EIBIS sample by comparing it against a series of randomly drawn samples from the relevant part of ORBIS. The idea of the latter is that in the absence of any selection bias, there should be no differences in observable characteristics between the two samples. Third, we compare the final EIBIS sample to two other databases: the Eurostat Structural Business Statistics (SBS) as well as the CompNet database. The aim is to assess how well information aggregated from the EIBIS survey represents dynamics in key variables and related cross country differences. The paper concludes with a summary of our findings.

1. What is the EIB Investment Survey (EIBIS)?[1]

EIBIS is an EU-wide survey that gathers qualitative and quantitative information on investment activities by non-financial corporates, both SMEs (with 5 to 250 employees) and larger corporates (with 250+ employees), their financing requirements and the difficulties they face. From 2019 onwards, a sample of firms from the United States (US) is included.

Using a stratified sampling methodology, EIBIS aims to be representative across all 27 Member States of the EU, the UK and the US, within countries, four firm size classes (micro,

[1] For more information on the EIBIS methodology – see https://www.eib.org/en/about/economic-research/surveys-data/about-eibis.htm

small, medium, and large) and four sector groupings (manufacturing, services, construction, and infrastructure). The survey is carried out through telephone (CATI) interviews in the local language.

All interviewed firms are drawn from the BvD ORBIS database which allows the survey answers to be linked to firms' financials and other administrative information.[2] The survey is annual with the first wave having taken place in 2016. Each year, the survey entails about 12,500 completed interviews with EU firms and more than 800 completed interviews with US firms. EIBIS is also designed to build a panel of observations over time, with roughly 40percent of firms from the previous wave being re-interviewed every year.

The main aim of EIBIS is to complement already available information on investment activities in the EU, and compare it – from 2019– with the US. It adds a firm-level dimension to existing macro-economic data on investment and thus allows for more fine-grained analyses of investment patterns (e.g. for different market segments). EIBIS also adds to firm-level investment data at the national level by providing full comparability of results across countries, as the same questions are asked to firms in all countries.

The distinctive feature of EIBIS vis-à-vis the Survey on the access to finance of enterprises (SAFE) of the European Commission and ECB (and similar initiatives) is its focus: in addition to asking few questions on access to finance, EIBIS collects detailed information on firms' investment activities and the link with investment financing decisions. The aggregate survey data, questionnaire, as well as a detailed account of the survey methodology, are available at www.eib.org/eibis.

2. What is the sampling frame used?

The EIBIS sampling frame for all countries is based on the BvD ORBIS dataset.[3] ORBIS is a commercial database, which contains data on 130 million firms worldwide; covering more than 100 countries.

ORBIS provides a list of firms by sector and country, a history of the firms' financials, as well as information on firms' directors, owners and patenting activities. The majority of information in ORBIS comes from business registers collected by local chambers of commerce to fulfill legal and administrative requirements. Bureau van Dijk organises the data and arranges them in a standard "global" format to facilitate company comparisons. For details, see Kalemli-Ozcan *et al.* (2015) or more recently Bajgar *et al.* (2020), who provide a detailed and comprehensive evaluation of the representativeness of ORBIS.[4]

[2] The matching is anonymised and EIBIS users can thus not identify individual firms.

[3] The ORBIS dataset that we rely upon is the so called 'historical dataset' which corresponds with the multi-disk dataset as described by Kalemli-Ozcan *et al.* (2015).

[4] Note that this section only assesses ORBIS as sampling frame. The representativeness of financials will depend on EIBIS information rather than those available from ORBIS.

The sub-population of interest for EIBIS is the non-financial corporate sector in the 27 EU Member States, the UK and the US, with at least five employees, belonging to one of the NACE categories C (manufacturing) to J (information and communication). Specifically, the EIBIS sampling frame is defined as all firms in ORBIS that:

- belong to the relevant size, sector and country groups
- are classified as 'industrial companies' in ORBIS (which means that they are not branches, inactive firms, public sector entities, financial firms, nor not-for-profit entities); and
- have recorded financials no older than three years.

For firms that report unconsolidated accounts and consolidated accounts, only unconsolidated entries are kept.

The sample includes a panel component as well as a top-up sample. Panel firms are those that participated in a previous wave of the survey, and that consented to be re-contacted in the following wave. The top-up sample, on the other hand, includes firms that did not participate in the preceding wave. The method adopted for selecting a top-up sample from ORBIS is random stratified sampling and the sample is stratified disproportionally by country, industry group and size class, and stratified proportionally by region within the country.

Weighting is done by calibrating the samples to Eurostat Structural Business Statistics (SBS) population data on the size/sector categories within each country for the EU. For the US, this was based upon several sources, namely the US Census Bureau data and the US Bureau of Economic Analysis (BEA) data.[5] Basing the weighting on population figures adjusts for any differences in the covered/uncovered firm profiles in addition to making corrections to the sample where it deviates from the quota profile due to non-response. Adjusting to the total population size means that the weights can be used for either single-country or cross-country analyses as the weighted samples reflect the correct proportions across countries.

3. How good is the sampling frame?

For EIBIS to be representative of the developments in the business sector it targets, a necessary condition is that the sampling frame from which survey respondents are drawn reflects as closely as possible the population of interest. In this section, we assess this criterion by benchmarking ORBIS against the Eurostat SBS[6]. The SBS data describe the structure of businesses across the European Union. The data are collected by National Statistical Institutes from statistical surveys, business registers, or various administrative

[5] For more information on how the US population data was constructed, please see the appendix to the Methodology report : https://www.eib.org/attachments/eibis-methodology-report-2019-en.pdf

[6] In specific comparisons we also use OECD Structural Business Statistics as it allows to include the US in our sample

... Continued from the previous page

		5-9	10-49	50-249	250+			5-9	10-49	50-249	250+
IE	Manufacturing	0.83	1.09	0.81	0.66	**IT**	Manufacturing	0.94	0.96	0.99	1.06
	Services	0.45	0.64	0.73	0.77		Services	1.02	1.03	1.04	1.01
	Construction	0.83	1.75	3.39	0.88		Construction	0.92	1.08	1.04	1.43
	Infrastructure	0.50	0.99	1.59	0.81		Infrastructure	0.80	1.06	1.00	1.03
LT	Manufacturing	0.67	0.81	0.90	1.00	**LU**	Manufacturing	0.43	0.79	1.37	1.32
	Services	0.81	0.90	0.97	1.12		Services	0.32	0.58	2.04	1.29
	Construction	0.60	0.90	0.84	0.81		Construction	0.35	0.58	1.59	1.39
	Infrastructure	1.02	0.96	0.94	1.10		Infrastructure	0.34	0.68	1.45	0.56
LV	Manufacturing	1.09	0.98	0.93	0.78	**MT***	Manufacturing	0.52	0.93	0.82	1.31
	Services	1.10	0.94	0.97	1.00		Services	0.28	0.79	0.78	3.00
	Construction	1.45	0.94	0.75	0.81		Construction	0.15	0.55	0.81	3.00
	Infrastructure	1.38	1.02	0.97	1.00		Infrastructure	0.50	0.82	0.98	1.67
NL	Manufacturing	0.53	0.89	0.68	0.97	**PL**	Manufacturing	1.28	0.08	0.13	0.27
	Services	0.62	0.89	0.67	0.69		Services	1.86	0.11	0.33	0.66
	Construction	0.49	0.91	0.77	0.69		Construction	3.31	0.17	0.34	0.62
	Infrastructure	0.57	1.00	0.73	0.71		Infrastructure	2.89	0.14	0.25	0.36
PT	Manufacturing	0.73	0.95	1.01	1.07	**RO**	Manufacturing	0.84	0.87	0.88	0.88
	Services	0.80	1.02	1.05	1.06		Services	0.84	0.91	1.01	1.05
	Construction	0.81	1.04	1.01	0.86		Construction	0.91	0.93	0.89	0.87
	Infrastructure	0.75	1.12	0.97	1.09		Infrastructure	0.90	0.99	0.96	1.05
SE	Manufacturing	0.64	0.92	0.89	1.00	**SI**	Manufacturing	0.79	0.98	0.93	0.97
	Services	0.81	0.90	0.84	0.89		Services	0.88	0.97	0.80	1.03
	Construction	0.74	1.05	1.05	1.21		Construction	0.74	0.92	0.87	1.10
	Infrastructure	0.58	1.00	0.94	1.09		Infrastructure	0.73	1.12	0.96	0.77
SK	Manufacturing	0.36	0.74	1.01	1.00	**UK**	Manufacturing	1.55	0.38	1.00	1.40
	Services	0.76	1.01	0.97	1.01		Services	1.31	0.65	1.09	1.32
	Construction	0.45	1.11	1.02	0.79		Construction	1.63	0.45	1.45	1.52
	Infrastructure	1.05	1.90	1.12	1.12		Infrastructure	1.18	0.72	1.53	1.81
US	Manufacturing	1.62	1.56	1.38	0.95						
	Services	1.01	0.85	0.79	0.64						
	Construction	1.24	1.45	1.53	1.16						
	Infrastructure	1.52	1.34	1.13	0.60						

Note: Coverage ratios for number of firms of EIBIS sampling frame compared to Eurostat SBS in 2019. Cells with coverage ratios >75 percent are highlighted in green. Cells with coverage ratios of 50<x<75 percent are highlighted in yellow. Cells with coverage ratios <50 percent are not highlighted. For all countries imputed cases were kept only if these were based on financial information that is no older than 3 years. For MT and CY this would have led to substantial under-coverage, so also cases were kept where financials were reported longer than 3 years ago. The calculations were carried out by IPSOS and this table is a replication of Table 5 of IPSOS (2019).

The reported coverage figures are broadly in line with those reported in Kalemli-Ozcan *et al.* (2015) or Bajgar *et al.* (2020). The better coverage in some instances can be explained by the fact that EIBIS limits itself to firms with 5 or more employees. Coverage in ORBIS improves with firm size as larger firms generally face stricter reporting requirements. As a result, excluding the smallest firms from the sample has a notable impact on coverage.

sources. The most pertinent variable in the SBS for our purposes is the number of enterprises by sector and size class in each country, which we can compare with the corresponding ORBIS data.

Table 1 below shows the coverage of our sampling frame (ORBIS) against Eurostat SBS (number of firms in ORBIS/number of firms reported in SBS).[7] In the majority of cases, coverage ratios are above 70 percent (highlighted in green); in 17 percent of the cases, they are between 50 percent and 75 percent (highlighted in yellow), while in only 11% of the cases coverage ratios are below 50 percent (no highlight).

Table 1 Coverage ratios of EIBIS sampling frame against EU SBS

		5-9	10-49	50-249	250+			5-9	10-49	50-249	250+
AT	Manufacturing	0.54	0.76	0.92	0.97	**BE**	Manufacturing	0.56	0.9	0.91	0.93
	Services	0.37	0.58	1.04	1.43		Services	0.52	0.88	0.94	1.04
	Construction	0.62	0.88	1.28	1.21		Construction	0.58	1.32	1.1	1.16
	Infrastructure	0.42	0.72	1.15	1.00		Infrastructure	0.50	1.14	1.09	1.04
BG	Manufacturing	1.22	1.13	1.05	1.05	**CY***	Manufacturing	0.15	0.40	0.86	1.67
	Services	1.3	1.4	1.27	1.25		Services	0.34	0.48	0.86	0.76
	Construction	1.32	1.48	1.17	0.73		Construction	0.16	0.27	0.61	7.00
	Infrastructure	1.64	1.49	1.35	1.35		Infrastructure	0.20	0.30	0.84	0.84
CZ	Manufacturing	0.25	0.69	0.81	1.03	**DE**	Manufacturing	0.66	0.79	0.86	0.81
	Services	0.35	0.80	1.32	1.07		Services	0.47	0.48	0.70	0.78
	Construction	0.25	0.73	1.15	1.06		Construction	0.67	0.75	0.98	0.97
	Infrastructure	0.35	0.88	1.14	1.11		Infrastructure	0.57	0.81	0.90	0.72
DK	Manufacturing	0.62	0.73	0.84	0.82	**EE**	Manufacturing	0.95	1.01	0.86	0.70
	Services	0.70	0.74	0.74	0.79		Services	1.06	1.23	0.86	0.80
	Construction	0.63	0.84	0.96	0.88		Construction	1.14	1.30	0.66	0.90
	Infrastructure	0.52	0.76	0.82	0.85		Infrastructure	0.99	1.20	0.83	0.77
GR	Manufacturing	0.27	0.74	0.88	0.97	**ES**	Manufacturing	0.64	0.86	0.92	0.95
	Services	0.16	0.18	0.60	1.19		Services	0.45	0.71	1.07	0.99
	Construction	0.15	0.12	0.29	0.60		Construction	0.62	1.01	1.64	0.93
	Infrastructure	0.19	0.34	0.56	0.84		Infrastructure	0.39	0.75	1.23	0.97
FI	Manufacturing	0.61	0.64	0.82	0.91	**FR**	Manufacturing	0.32	0.43	0.69	0.92
	Services	0.81	0.69	0.92	0.92		Services	0.59	0.51	0.80	0.85
	Construction	0.70	0.69	0.94	1.03		Construction	0.29	0.40	0.95	1.29
	Infrastructure	0.83	0.74	1.19	0.83		Infrastructure	0.30	0.48	0.77	0.95
HR	Manufacturing	0.68	0.91	0.89	1.03	**HU**	Manufacturing	0.71	1.00	1.03	1.08
	Services	0.78	0.86	0.87	0.95		Services	0.83	1.06	1.07	1.08
	Construction	0.95	0.95	0.91	0.50		Construction	0.80	1.11	0.98	0.75
	Infrastructure	0.86	1.01	0.98	1.07		Infrastructure	0.66	1.14	1.01	0.96

... Continued on next page.

[7] The Eurostat SBS only report size class 0-9 employees. To divide this into 0-4 and 5-9 employee size classes, we fitted a Zipf distribution and derived the corresponding values this way.

3. Is selection bias an issue?

Having a good sampling frame is a necessary condition to select a representative sample and thus for the EIBIS survey results to reflect developments in the non-financial corporate sector of Europe and the US in terms of investment activities. It is, however, not sufficient and the firms that eventually participate in the survey may suffer from selection bias.

One way to assess the quality of our sample is by comparing itwith randomly drawn samples from the sampling frame. EIBIS is a telephone survey based on a probability sampling approach with a target quota.[8] There is nevertheless still a risk, at least in theory, that interviewers end up calling only firms that are easy to get hold of, and thereby introducing a type of convenience element to sampling. Any issues that would imply a non randomly selected subset of firms from the sampling frame may seriously bias the data.

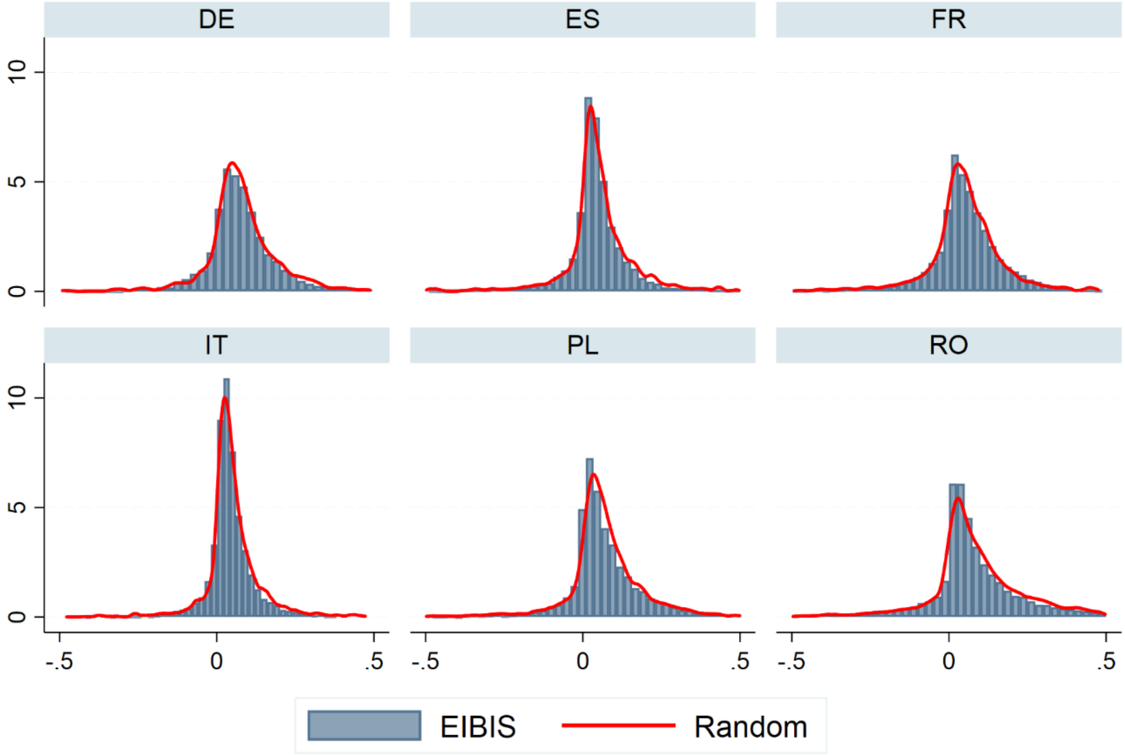
To test whether, and if so to what extent, the EIBIS sample is subject to such selection issues, we compare the unweighted distribution of a set of ORBIS variables in the final EIBIS sample with five randomly drawn samples from the same sampling frame. The rationale is that in the absence of any selection bias, we should find that the distribution of these variables is statistically identical in the EIBIS sample and random samples. In order to keep our methodology aligned with the stratas in the survey sample, each random sample comprises of random draws with similar sample size as in the EIBIS sample for each of the four sectors and four firm-size groups. We compare the different samples based upon a set of financial variables with sufficient coverage in ORBIS, namely sales growth, cash flow, cash holding, equity ratio, investment rate, return on assets, trade debt, trade credit and leverage. Detailed definitions of the different variables used are shown in the Appendix.

In a first step, we visually compare our EIBIS sample with the random samples drawn from Orbis. We observe that distributions overlap which suggests sufficient randomness of the EIBIS sample for the overswhelming share of variables across countries.[9] As an example, Figure 1 plots the return on assets distribution from EIBIS by country as well as the distribution of the same variable from a random sample of firms from the Orbis sampling frame.

[8] This means that interviewers have a pool of 8 firms as overall target quota rate to achieve a completed interview. Note that realised quota rates vary across countries with some countries achieving better response rates, and some slightly worse ones.

[9] For brevity we do not include all graphs here. They are available from the authors on request.

Figure 1. Return on assets in EIBIS survey vs random sample from Orbis (selected countries)



Note: the figure shows the distribution of return on assets for the EIBIS sample firms in 2017 (red solid) against the distribution of the same variable from a random sample of the corresponding Orbis sampling frame (blue dashed). For easier readability, the range of the x-axis is limted to -0.5 to 0.5.[10]

In the next step, we test the similarity of the variable distributions from the EIBIS and random samples using two-sample Kolmogorov-Smirnoff (KS) tests. To facilitate a more meaningful comparison and allow for a greater power of the test, we run the analyses only on variables that have at least 200 non-missing values in both the EIBIS sample and the random sample for a country in a given year.

Table 2 gives an overview of the average success rate of the KS-tests for the different variables, across all countries. A KS test is considered successful if the EIBIS sample is not statistically different from a random sample drawn from Orbis. For the majority of variables and countries, the success rate is very high suggesting the EIBIS sample is very similar to the random sample. For more than 90 percent of the country-variable cases the success rate is above 0.75. Ther are 4 percent of the cases when the rate is below half. [11]

[10] Note that firms included in the 2019 survey are asked about the 2018 or earlier performace, while Orbis variables are often only available for them from year 2017.

[11] In a couple of instances, the KS test rejects the null hypothesis that the distributions come from the same underlying population. However, this only happens for variables for which the cdf is not continuous, such as investment ratio and trade credit, for which KS tests are not the ideal set-up. The former variable has a highly skewed distribution driven by the lumpy nature of investment while trade credit has a distribution censored at 0. In addition, some of the weaker test results are highly driven by individual sectors. The weaker results for trade credit in Poland for example originate from the infrastracture sector, and in Germany from the construction sector.

Table 2: Results from comparing EIBIS to random draws from the sampling frame

	Sales growth	Cash flow ratio	Cash holding	Equity ratio	Investment ratio	Leverage	Return on Assets	Trade credit	Trade debt
AT	1.00	1.00	1.00	n.a.	1.00	n.a.	n.a.	n.a.	1.00
BE	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00
BG	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
HR	0.93	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CZ	1.00	1.00	1.00	0.93	0.80	1.00	1.00	0.93	1.00
DK	n.a.	1.00	0.93	0.93	1.00	1.00	1.00	1.00	1.00
EE	1.00	1.00	1.00	1.00	0.73	1.00	0.73	1.00	1.00
EL	1.00	0.93	1.00	1.00	1.00	1.00	1.00	0.87	0.93
FI	1.00	1.00	0.87	0.93	1.00	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00
DE	1.00	n.a.	1.00	n.a.	1.00	n.a.	1.00	0.27	0.93
HU	0.87	0.93	1.00	0.87	0.40	0.93	0.67	1.00	1.00
IE	n.a.	n.a.	0.93	n.a.	0.40	1.00	0.67	1.00	0.27
IT	1.00	1.00	0.93	0.87	0.67	0.93	0.93	1.00	1.00
LV	1.00	n.a.	1.00	0.93	0.93	n.a.	1.00	1.00	1.00
LT	1.00	n.a.	1.00	1.00	1.00	n.a.	n.a.	n.a.	n.a.
NL	n.a.	n.a.	1.00	n.a.	1.00	n.a.	n.a.	n.a.	1.00
PL	1.00	1.00	1.00	0.47	0.33	n.a.	1.00	0.53	0.47
PT	0.67	1.00	1.00	1.00	0.93	0.87	1.00	1.00	1.00
RO	0.87	0.80	1.00	0.93	1.00	0.93	n.a.	n.a.	1.00
SK	0.93	1.00	1.00	0.93	1.00	0.93	1.00	0.93	1.00
SI	1.00	0.93	1.00	0.93	1.00	0.87	1.00	1.00	1.00
ES	0.93	1.00	0.93	1.00	0.80	0.93	0.87	1.00	1.00
SE	0.93	0.87	0.93	0.73	1.00	1.00	1.00	0.93	1.00
UK	1.00	1.00	1.00	1.00	0.73	1.00	1.00	0.33	0.27

Note: Each cell reports the average share of KS tests that find that distributions are similar for a given variable and country in EIBIS and a random sample drawn from Orbis, over 3 sample periods and 5 random draws. For example, the 1.00 for sales growth in Austria means that 100 percent of the KS tests find that sales growth in the EIBIS sample is equivalent to a random sample. The n.a. stands for non-available, i.e. the variable does not provide a sample of 200 non-missing observations. Countries where all variables are n.a are not shown in the table: CY, LU, MT and US.

The EIBIS survey has as an important panel component. The creation of a panel could by itself compromise data quality as this may again result in a selection bias of 'surviving' firms or better-performing firms. An important question is thus whether the effort to create a panel structure compromises data quality. In order to assess it, we repeat the same KS testing procedure as described above, only for the panel firms in the EIBIS sample. The results, shown in Table 4 in the Appendix, indicate that the panel-bias is not a cause for concern.

An alternative approach to test the similarity of the EIBIS sample with randomly drawn samples is to perform *t*-tests on the mean. Given the less stringent requirements on the full distribution, *t*-tests can be performed at a fine-grained level, by comparing sample averages within each sector-size group, resulting in 16 tests for each random draw.[12] The results lead to the same conclusion as KS tests, as suggested by Table 5 in the Appendix.

4. Aggregate statistics based on the EIBIS sample – a comparison with other data.

Another way to assess the quality of our sample is to compare aggregate statistics based on firms in the EIBIS sample to aggregate statistics in other data sources. To do so, we can use both SBS and CompNet, as outlined below.

When comparing to SBS we focus on cross-section, and compare average statistics over the 2016-2017. This is the latest period available in the sampling frame. We compare average size and labour productivity across countries. In order to run a meaningful comparison on aggregate statistics, we apply value-added weights to compare productivity and population weights to capture average firm size.

Whe comparing to statistics from CompNet we look at four year periods 2013-2016 or 2010-2013 when CompNet data is only available from earlier vintages. We look at the variables listed in Table 2. We compare EIBIS to unweighted CompNet statistics on median ratios thus we apply population weight in EIBIS.

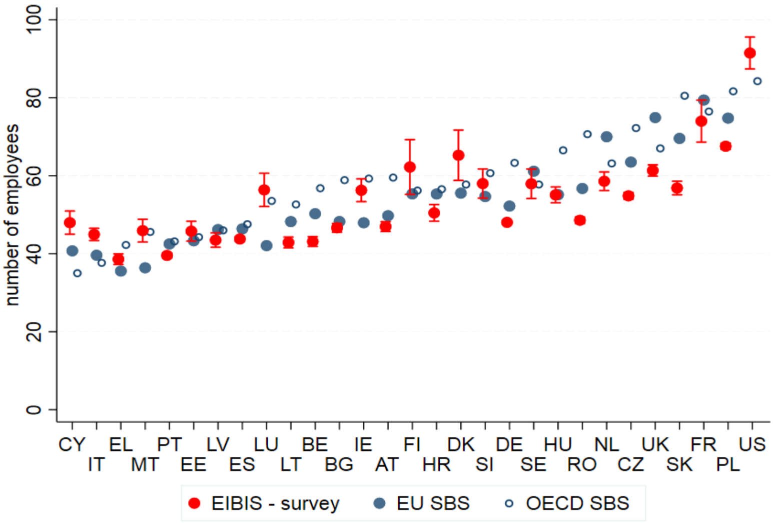
4.1 Comparison to SBS

Apart from assessing the coverage of our sampling frame, SBS data allow us to cross-validate the EIBIS survey results.

Figure 2, for example, shows the average firm size of the EIBIS survey compared to the average firm size for the corresponding population of firms (sector and size-class range) from EU SBS. The figure also includes the comparable OECD statistics to facilitate comparison with US data. We find that the representative firm size statistics calculated from EIBIS are very close to the Eurostat and/or OECD population averages.

[12] In addition to the 200 non-missing values, we require a minimum of 10 non-missing observations within a firm-size sector group to run the *t*-test

Figure 2 Comparison of average firm size: EIBIS, Eurostat SBS and OECD SBS

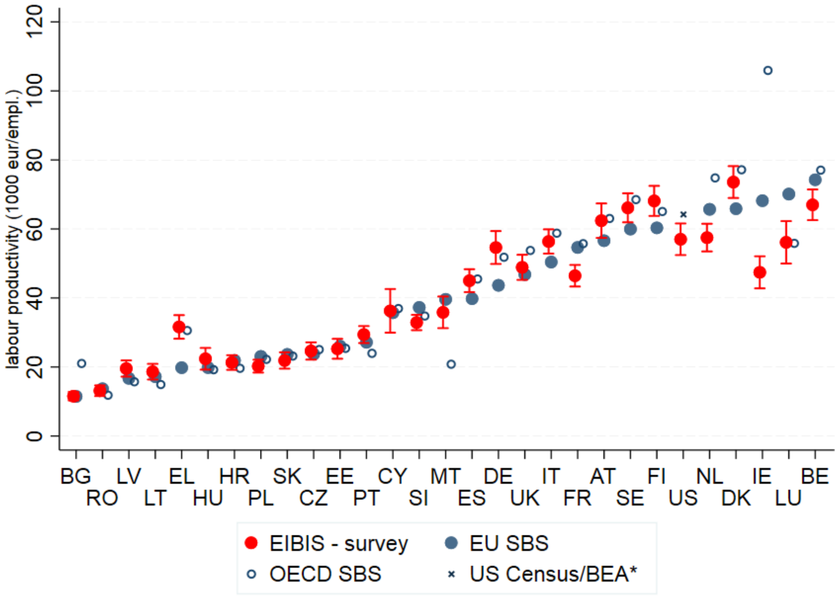


Note: Firm size average statistics of EIBIS survey compared to Eurostat and OECD SBS sources. SBS averages are calculated as population-weighted averages from size-class and sector averages for firms above 10 employees. The values are averaged over two years, 2016 and 2017. EIBIS figures (in red) also display the range of the sample mean estimate: plus and minus 1.96 times the standard deviation from the mean.

Next to firm size, we can also assess how labour productivity based on the EIBIS survey compares to population statistics. Figure 3 shows the average labour productivity of EIBIS survey firms compared to the average productivity of the corresponding population of firms (sector and size-class range) from EU SBS. We define labour productivity as nominal value added in EUR over the number of employees of the firm. Besides, we also add labour productivity estimates calculated from OECD structural business statistics and US Census data for the US.

We find that average labour productivity statistics from the EIBIS survey highly overlap with official statistics. Large discrepancies are observed in the case of one country only: Ireland. In both cases, the value-added weight in EIBIS is capped due to the highly skewed nature of the underlying value-added distribution.

Figure 3 Comparison of average firm productivity: EIBIS, Eurostat SBS, OCED SBS and US Census/BEA



Note: Labour productivity (value-added over firm size) average statistics of EIBIS survey compared to Eurostat and OECD SBS sources. For the survey statistics, value-added weight is employed. SBS averages are calculated as population-weighted averages from size-class and sector averages for firms above 10 employees. The values are averaged over two years, 2016 and 2017. EIBIS figures (in red) also display the range of the sample mean estimate: plus and minus 1.96 times the standard deviation from the mean.

4.2 Comparison to CompNet

Next, we compare our sample to the CompNet database. CompNet data is based on a "distributed micro-data approach": that is, relevant data are extracted from – often confidential – firm-level datasets available within National Central Banks or National Statistical Institutes and aggregated such that the confidentiality of firm data is preserved. The outcome is a wide range of financial indicators, based on micro-level data for 12 EU countries. To assess the quality of the EIBIS sample, we re-produce the same indicators using the ORBIS database and compare them with those in CompNet.

We first need to note that CompNet data exists for 'all firms' and for firms with '20 or more employees'. Insofar as EIBIS excludes the smallest firms (0-4 employees) but includes firms with between 5 and 19 employees, this raises the question of which CompNet dataset to use for our benchmarking exercise.

We opted for the '20+' database. One of the main reasons is that firms between 0 and 4 employees account for a disproportionately large share of the number of firms in the business economy (much larger than firms between 5 and 19 employees). Including them will put the focus of the comparison on a segment that is excluded from EIBIS; especially if population weights are used for this comparison as these give a lot of weight to the smallest firms in the population.[13] Another reason to use the 20+ CompNet is rooted in the fact that financial information is less frequently available for smaller firms in ORBIS.

In addition, we have to bear in mind that, even after having selected the most comparable CompNet data, a full comparison between ORBIS and CompNet is not possible for all countries. Some variables in ORBIS are less well filled in some countries. For example, while EBITDA has high coverage for most countries, this is not the case in the Netherlands, Austria and Germany. Hence, for these countries, a comparison between ORBIS and CompNet is difficult and a mismatch in statistics does not necessarily imply that the EIBIS sample is inadequate. Similarly, across the different vintages of CompNet, there are differences in country coverage and the range of available financial statistics

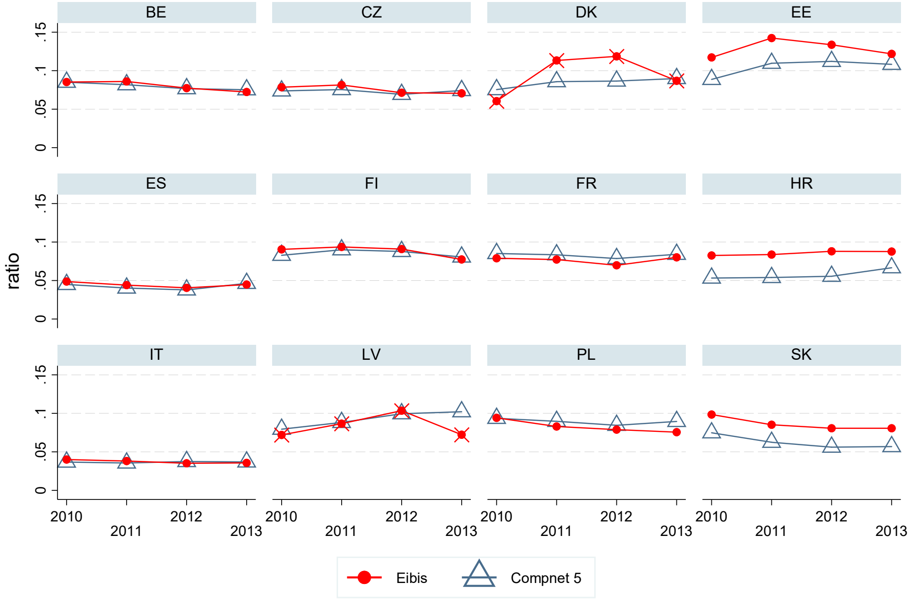
In this report, we use the three most recent vintages of CompNet, namely vintages 5, 6 and 7 in order to get the highest country and variable coverage. We compare the EIBIS sample and CompNet based on the median of the following variables: cash flow ratio, cash holding, equity ratio, investment ratio, leverage, return of assets (ROA), trade credit ratio, trade debt ratio.[14]

Figure 4 plots the "cash flows to total assets ratio" as reported both in ORBIS (for the firms in the EIBIS sample) and the CompNet 20+ data. It suggests a very close match of the median for this variable between the two sources in seven countries. In the remaining countries, we only observe minor discrepancies. Some of the deviations may result from a lower coverage in ORBIS, which is the case in Denmark and Latvia. In these latter countries, the ORBIS variable is only available for less than 50 percent of the firms, as shown by the cross on the data points in Figure 4.

[13] That is because CompNet data is weighted by population weights.

[14] The definitions of the different variables, in line with those used in CompNet, is presented in the Appendix.

Figure 4 Cash flow ratio: EIBIS vs CompNet



Note: Comparison of median in CompNet 20+ dataset and EIBIS sample for the period 2010-2013; weighted by population weights. In this comparison, only CompNet vintage 5 is used because this variable is not produced for later vintages. The Figure only includes the countries where cash-flow ratio is calculated in CompNet: Belgium, Czechia, Denmark, Estonia, Spain, Finland, France, Croatia, Italy, Latvia, Poland, Slovakia. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms only due to variable availability.

A similar pattern as for the cash flows to total assets ratio holds for most variables that we consider. A complete overview is included in Figures 5-12 in the Appendix. Table 4 offers a summary of this comparison for all variables considered. A value of 1 means that the series from the two sources are similar; and 0 that they are not.[15]

The overall positive results of our comparison between CompNet and the ORBIS variables for the EIBIS sample is in line with the findings reported in Ferrando *et al.* (2015), where the 'all firms' CompNet database is compared with ORBIS. The authors conclude that "most of [the indicators] share similar dynamics among the two datasets, except for some minor discrepancies".

[15] We also marked comparisons, where Orbis values are missing for a significant share of observation (*), or discrepancies between the two datasets are limited to a single period (^{m}).

Table 3. Comparison of median in EIBIS and CompNet

	Sales growth	Cash flow ratio	Cash holding	Equity ratio	Investment ratio	Leverage	Return on Assets	Trade credit	Trade debt	Total
BE	1	1	1	1	0	1	1	1	1	8/9
CZ	1	1	0	1	1*	1	1	1	1	8/9
DE	na	na	na	na	na	1	1^m	1	0^m	3/4
DK	1	1^m	1^m	1	0^m	1	1^m	1^m	1	8/9
EE	1*	1	1	1	1	na	1	0	1	7/8
ES	1	1	1	1	1	1	1	1	1	9/9
FI	1	1	1	1	1*	1	1	1	1	9/9
FR	1	1	1*	1	1	1	1	1	0	8/9
HR	1*	1	1	1	1	1	1	1	1	9/9
HU	na	na	1*	na	1*	1*	1*	na	1	5/5
IT	1	1	1*	1	1*	1	1	1	1	9/9
LT	na	na	1	0	na	1	0	0^m	1	3/6
LV	0	1	1	0	1	1	0	1	1	6/9
NL	na	na	0	na	0	0^m	1^m	0^m	0	1/6
PL	1*	1	1*	1*	1	1*	1	0*	1*	8/9
PT	na	na	1	na	1	1	1	1	1	6/6
RO	na	na	0	na	1	0^m	1	1	1	4/6
SE	na	na	1	1	0	1	0	1	1	5/7
SI	na	na	1	1	1	1	1	1	1	7/7
SK	1	1	na	1	1	0	0	1	1	6/8
Total	12/12	12/12	15/18	13/15	15/18	16/19	16/20	15/19	17/20	

Note: Comparison of CompNet 20+ dataset and EIBIS sampling frame for the period 2006-2013; weighted by population weights. Qualitative assessment. 1 means that the two samples are similar; 0 otherwise. *= comparison valid except for 1 year (out of 4 years), where there is a large difference. ^{m}= the ORBIS variable is missing for more than half of the firms in some years. na=not available. Total: number of cases where the samples are similar.

Conclusion

This paper reviewed the quality of the EIBIS data by:

1. assessing the adequacy of its sampling frame (the Bureau van Dijk ORBIS database);
2. assessing the degree to which the final EIBIS sample differs from a series of randomly drawn samples from the same sampling frame;

3. examining selection bias of the firms included in the sample by comparing indicators for those firms with Structural Business Statistics (SBS) data and CompNet.

All tests yield largely positive results: the sub-sample of ORBIS that acts as sampling frame for all countries in EIBIS shows a good match when benchmarked against the Eurostat SBS and the OECD SBS.

Comparing the final EIBIS sample with 5 random samples from the same sampling frame shows that there is little evidence for sampling bias in EIBIS. The latter could in principle have arisen due to systematic differences in willingness to participate in the survey across firms; the quota dimension of EIBIS and/or the panel component in the data. Comparing the distributions of a series of financial variables between the final EIBIS sample and several random samples from the same sampling frame, we find, with very minor exceptions, no evidence for this.

The positive coverage of the EIBIS sample is also reflected in a comparison of the aggregate financial information contained in ORBIS with that in Eurostat SBS, OECD SBS and CompNet data. Despite some difficulties in making the data sources comparable, we find a high degree of overlap in the evolution of a series of financial indicators in the two data sources.

Overall, the evidence reported in this paper suggests that EIBIS provides a representative dataset to study the investment situation in the EU corporate sector.

References

Bajgar, Matej, Giuseppe Berlingieri, Sara Calligaris, Chiara Criscuolo and Jonathan Timmis (2020). "Coverage and representativeness of Orbis data." OECD Science, Technology and Industry Working Papers, No. 2020/06.

Ferrando, Annalisa, Matteo Iudice, Carlo Altomonte, Sven Blank, Marie-Hélène Felt, Philipp Meinen, Katja Neugebauer and Iulia Siedschlag (2015). "Assessing the financial and financing conditions of firms in Europe: the financial module in CompNet." ECB Working Paper No 1836.

Ipsos (2019) "EIB Group Survey on Investment and Investment Finance" Technical report, Ipsos MORI Social Research Institute
https://www.eib.org/attachments/eibis-methodology-report-2019-en.pdf

Kalemli-Ozcan, Sebnem, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych and Sevcan Yesiltas (2015). “How to Construct Nationally Representative Firm Level data from the ORBIS Global Database”. NBER Working Paper No. 21558

Appendix A – Data cleaning procedures

EIBIS – ORBIS comparison

To ensure the comparability of the EIBIS-ORBIS matched database with the random samples extracted from the ORBIS population, we apply to the latter the same cleaning procedure used in the former.

First, the procedure replaces with missing values the negative values of variables coming from Balance Sheet data, excluding those that could have a negative value such as Shareholders Fund or Current Net Assets.

Second, we annualise "flow variables" coming from a firm's Income Statement: variables of firms adopting a financial period different from 12 months are harmonized to 12 months.

Third, the routine creates a new variable for loans. In case of missing values for loans and value 0 for Current Liabilities, the loans variable takes value 0. Then, if Current Liabilities are set at 0 and creditors is missing, the latter takes a value 0. Afterwards, the routine replaces loans missing values with the resulting value coming from the subtraction of Trade Payables (creditors) and Other Current Liabilities to Current Liabilities. If Other Current Liabilities are missing, only the subtraction of Payables to Current Liabilities is used. Finally, negative values are set to missing.

After these small adjustments, an ad-hoc cleaning is performed on ratios and indicators. Values that are considered as outliers basing on thresholds defined by looking at the data distribution and common values of each ratio within economic literature are dropped.

This cleaning procedure never sets to missing more than 1.6 per cent of the observations, with an average percentage of set-to-missing values of 0.25 per cent and a median of 0.16 per cent (for EIBIS-ORBIS matched data). These percentages show how this procedure is still a conservative one, dropping only outlier values.

After the ad-hoc cleaning, winsorizing at the 1st and 99th percentile at the Country-Sector-Year level is performed.

Variables	Definition
Sales growth	Turnover / Turnover (t-1)
Cash flow ratio	Cash flow / Total assets
Cash holding	Cash & Equivalents / Total assets
Equity ratio	Shareholder's fund / Total assets
Investment ratio	(Change in fixed assets + depreciation) / Fixed assets (t-1)
Leverage definition 1	Short and long term debt / Total assets
Leverage definition 2	(Current and non-current liabilities - creditors) / Total assets
Return on assets	Operating profit / Total assets
Trade credit	Creditors / Total assets
Trade debt	Debtors / Total assets

EIBIS – COMPNET comparison

To ensure the comparability of EIBIS-ORBIS matched database with COMPNET, we adopt the same cleaning procedure used in the latter and described in details in Ferrando et al. (2015).

We build the same ratios as described in their paper, and perform the two-step outlier treatment used for COMPNET data. The first step trims the distribution of each indicator by country-sector at the 1st and 99th percentile. The second step eliminates values that fall outside the interval determined by the median of the distribution of each indicator (by country-sector) plus/minus ten times the interquartile range of the same distribution.

Appendix B – Additional results

Table 4. Results from comparing EIBIS PANEL to random draws: KS tests

	Sales growth	Cash flow ratio	Cash holding	Equity ratio	Investment ratio	Leverage	Return on Assets	Trade credit	Trade debt
AT	1.00	1.00	1.00	1.00	n.a.	n.a.	n.a.	n.a.	1.00
BE	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00
BG	1.00	1.00	0.87	1.00	1.00	1.00	1.00	1.00	1.00
HR	1.00	0.93	1.00	0.93	1.00	1.00	1.00	1.00	1.00
CZ	1.00	1.00	1.00	0.67	1.00	0.87	1.00	1.00	1.00
DK	n.a.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EE	0.93	1.00	1.00	0.73	1.00	0.87	1.00	1.00	1.00
EL	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.93
FI	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00	1.00	0.73	1.00	1.00	1.00
DE	1.00	n.a.	1.00	1.00	n.a.	1.00	n.a.	0.27	1.00
HU	0.93	1.00	1.00	0.67	0.80	0.87	0.93	0.93	1.00
IE	n.a.	n.a.	1.00	0.40	1.00	0.60	n.a.	1.00	0.27
IT	0.93	1.00	0.93	0.47	0.93	1.00	1.00	1.00	1.00
LV	1.00	n.a.	1.00	0.93	n.a.	1.00	0.93	0.93	1.00
LT	1.00	n.a.	1.00	0.90	n.a.	n.a.	1.00	n.a.	n.a.
NL	n.a.	n.a.	1.00	1.00	n.a.	n.a.	n.a.	n.a.	1.00
PL	1.00	1.00	1.00	0.33	n.a.	1.00	0.47	0.27	0.40
PT	0.67	1.00	1.00	0.93	0.93	0.93	1.00	1.00	1.00
RO	1.00	0.73	1.00	1.00	0.73	n.a.	0.93	n.a.	1.00
SK	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SI	1.00	0.93	1.00	1.00	0.80	1.00	0.80	1.00	1.00
ES	0.93	1.00	0.93	0.93	1.00	0.93	1.00	1.00	1.00
SE	1.00	0.67	0.93	1.00	1.00	1.00	0.73	1.00	1.00
UK	1.00	1.00	1.00	0.93	1.00	1.00	1.00	0.33	0.33

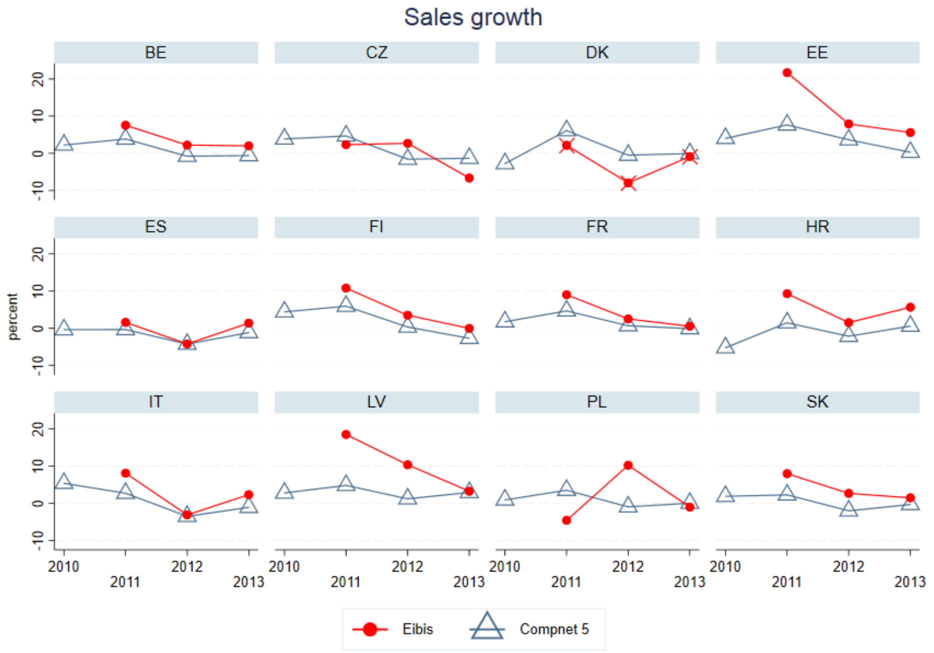
Note: Each cell shows the average share of Kolmogorov-Smirnoff (KS) tests that find that distributions are similar over 3 sample periods and 5 random draws for a given variable and country. For example, the 1.00 for the sales growth in Austria means that 100 percent of the KS tests suggest that sales growth in the EIBIS sample is equivalent to a random sample. The n.a. stands for non-available, i.e. the variable does not provide a sample of 200 non-missing observations. Countries where all variables are n.a are not shown in the table.

Table 5. Results from comparing EIBIS PANEL to random draws: *t*-tests

	sales growth	cash flow	cash	equity	invest-ment rate	ROA	Leverage	trade credit	trade debt
AT	97	97	95	95	97	96	96	53	96
BE	93	94	92	93	98	94	90	95	85
BG	88	98	93	95	94	95	95	97	98
HR	94	88	97	96	95	89	99	92	93
CY*	100		100	100		100	100	100	100
CZ	89	93	97	77	98	95	84	90	96
DK	98	93	93	97	98	95	94	97	94
EE	95	93	98	89	95	96	91	99	93
EL	92	98	87	93	100	99	94	97	95
FI	93	96	96	94	95	96	96	98	92
FR	95	91	97	84	98	95	90	92	95
DE	95	98	88	90	99	95	92	83	95
HU	89	91	87	88	93	91	87	95	93
IE			92	74	96	67	82	78	76
IT	84	96	83	83	93	98	84	95	70
LV	92		89	92		93	93	97	92
LT	92		99	93		96	94	98	99
LU*	91	97	95	100	98	97	100	99	95
MT*	100		100	100		85	89	96	95
NL	33	63	97	98	97	97			86
PL	93	89	92	77	62	89	90	73	76
PT	88	97	88	90	93	97	92	95	90
RO	87	89	90	93	93	88	85	91	84
SK	95	96	94	96	93	97	97	92	98
SI	95	96	91	99	94	97	100	86	93
ES	93	94	89	85	93	96	88	95	87
SE	95	97	97	94	95	95	94	93	90
UK	96	100	94	93	97	96	90	76	65
US*	100	100	100	80	100	100	73	100	100

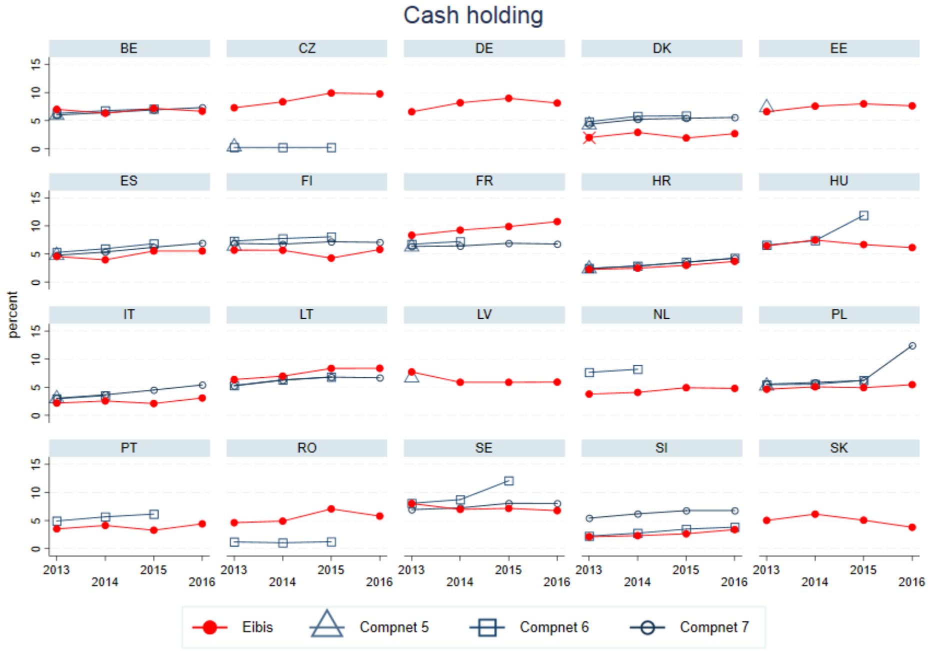
Note: Each cell shows the share of *t*-tests, where equality of variable means cannot be rejected, averaged over three years and the sizeclass-sector groups, in the EIBIS sample and random draws. For example, 97 percent for sales growth in Austria means that the overwhelming majority of the *t*-tests (12 sizeclass-sector groups, over 5 random draws and 3 years) suggest that the average of sales growth in the EIBIS sample is similar to a random sample.* For these countries, the 200 non-missing variable restrictions were not applied.

Figure 5 Sales growth: EIBIS vs CompNet



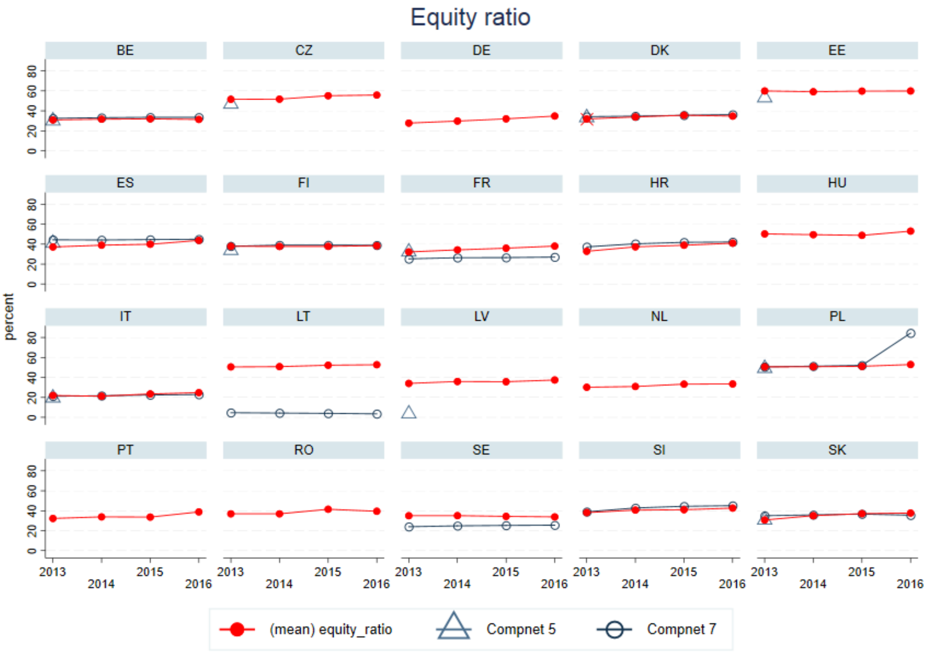
Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2010-2013; weighted by population weights. The figure includes all countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 6 Cash holding: EIBIS vs CompNet



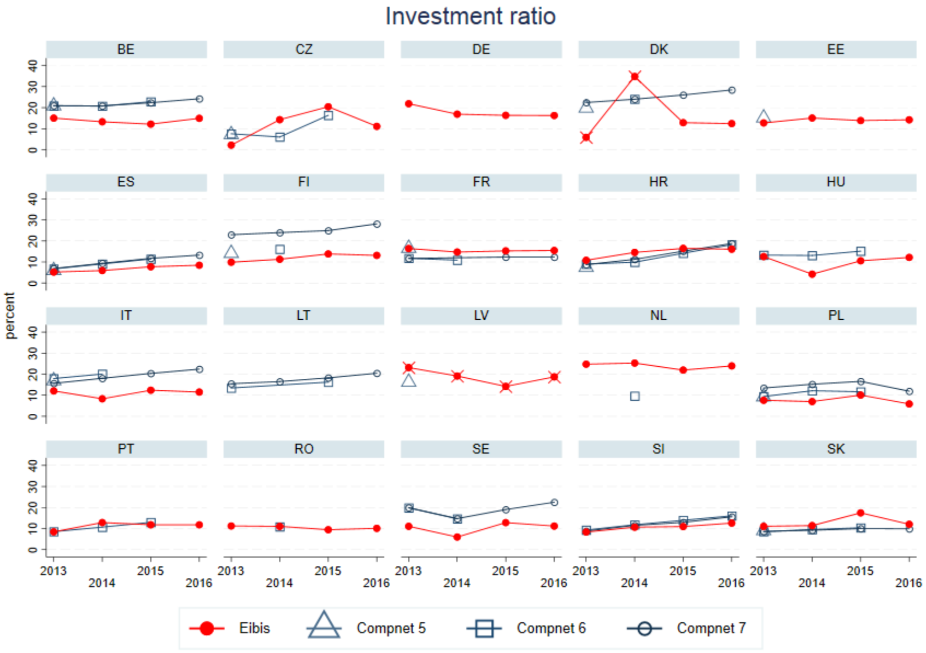
Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figure includes all countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 7 Equity ratio: EIBIS vs CompNet



Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figures inlcudes all countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

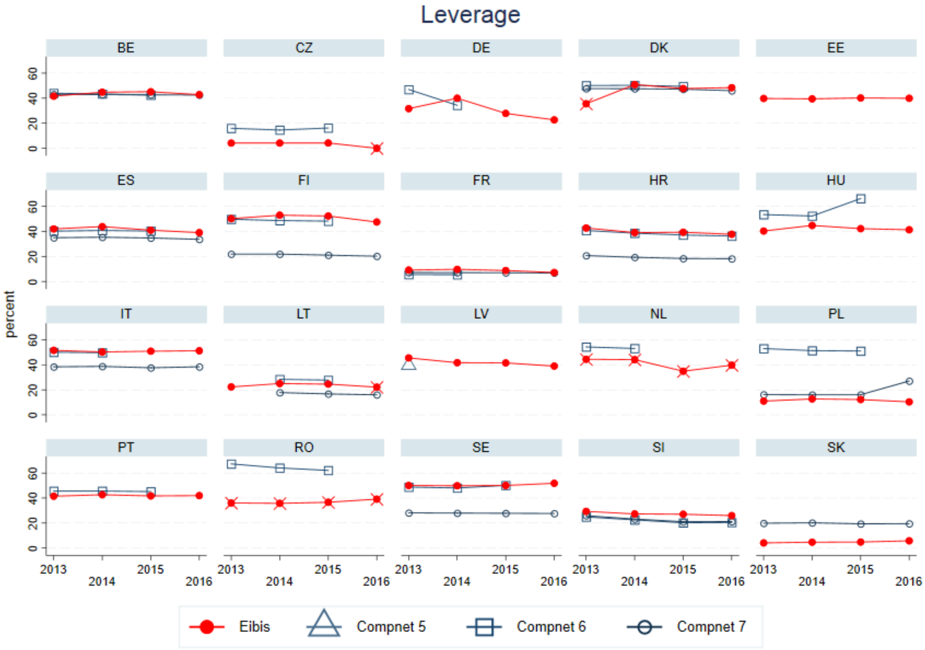
Figure 8 Investment ratio: EIBIS vs CompNet



Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figure includes all countries that participated in any of the three

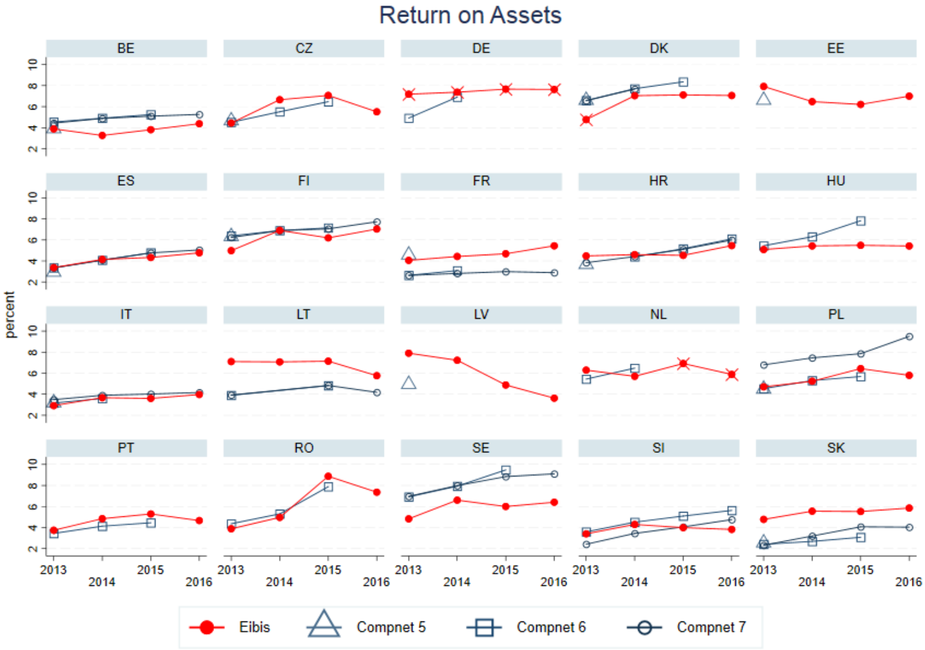
CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 9 Leverage: EIBIS vs CompNet



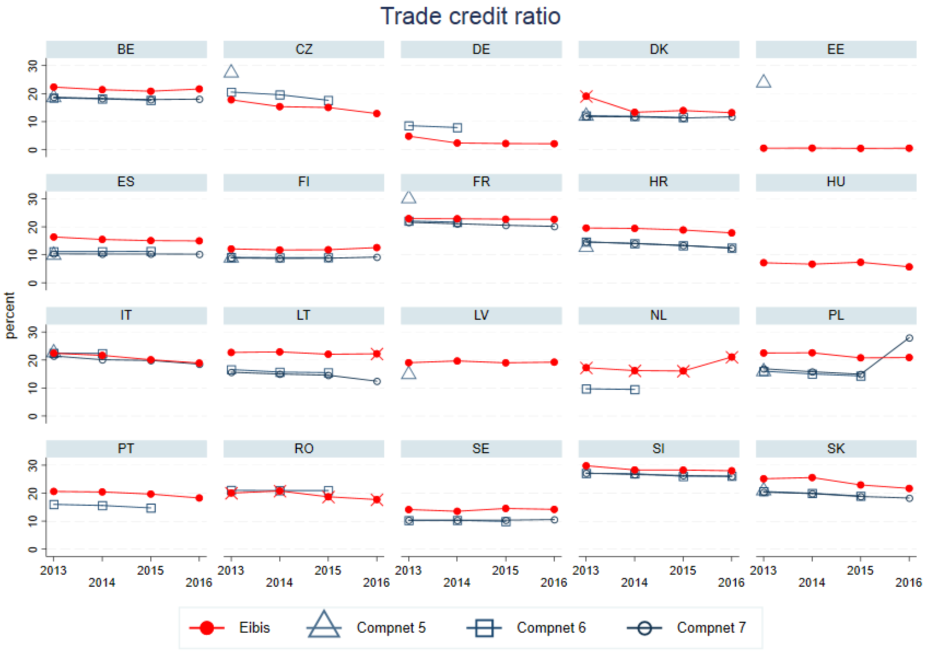
Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figure includes all countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 10 Roa: EIBIS vs CompNet



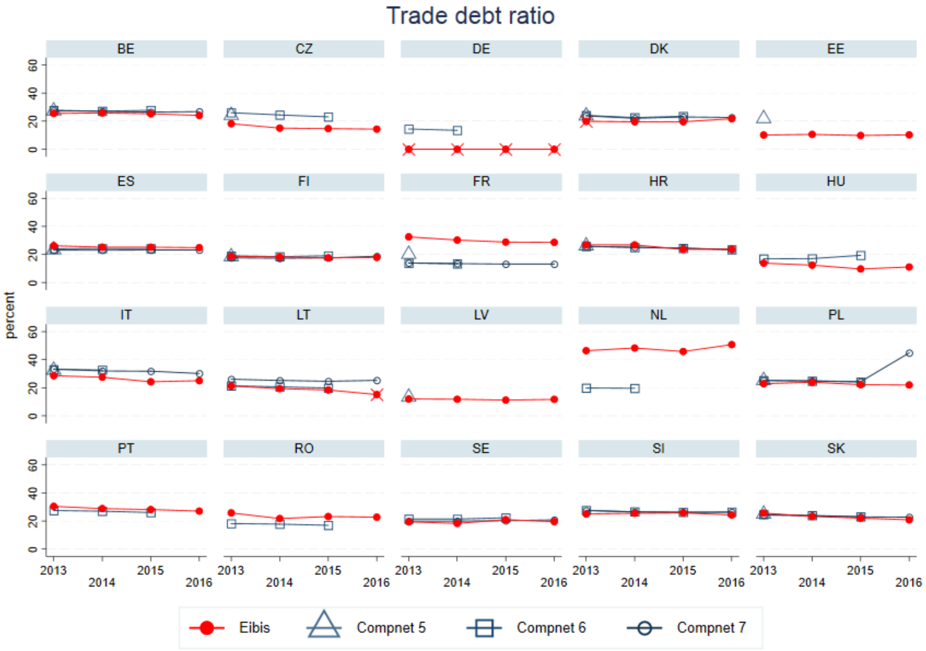
Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figure includes all countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 11 Trade credit: EIBIS vs CompNet



Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The Figure includes All countries that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

Figure 12 Trade debt: EIBIS vs CompNet



Note: Comparison of median in CompNet 20+ dataset and EIBIS sampling frame for the period 2013-2016; weighted by population weights. The figure includes all countries displayed that participated in any of the three CompNet vintages. EIBIS statistics are marked with a cross when the underlying statistic is calculated from less than half of the firms due to variable availability.

ECONOMICS – WORKING PAPERS 2020/08

EIB Group Survey on Investment and Investment Finance

A technical note on data quality

European
Investment
Bank
The EU bank

Economics Department
✉ **economics@eib.org**
www.eib.org/economics

European Investment Bank
98 -100, boulevard Konrad Adenauer
L-2950 Luxembourg
☎ +352 4379-22000
www.eib.org – ✉ **info@eib.org**

PDF: ISBN 978-92-861-4718-0