An Analysis of Business Response Burden and Response Behaviour Using a Register of Data Provision

Boris Lorenc[1], Wim Kloek[2], Linda Abrahamsson[3], Stephanie Eckman[4]

[1]Statistics Sweden, e-mail: boris.lorenc@scb.se

[2]Eurostat, e-mail: wilhelmus.kloek@ec.europa.eu

[3]Karolinska Institutet, e-mail: linda.abrahamsson@ki.se

[4]Institute for Employment Research, e-mail: Stephanie.Eckman@iab.de

Abstract

An analysis is carried out of the total burden that a national statistical institute imposes on businesses for the purpose of data collection, as well as of the businesses' response behaviour. A descriptive overview presents distribution of the response burden, expressed in terms of number of surveys, survey instances, and total expected time under compliance, among business size and industry classes. A multi-level analysis indicates that too high levels of response burden of a business lead to business's reduced participation in new survey instances. There is an indication that businesses that are nonrespondents in the first instance of a survey risk remaining nonrespondents, indicating that a particular attention needs to be paid to them.

1. Introduction

In a thought-provoking article, McCarthy et al. (2006) examined in the context of establishment surveys the relationship between prior reporting burden placed on sampled units in several agricultural surveys and the response pattern of these same units in later surveys. Their results indicated that burden − measured in terms of the number of other surveys these agricultural operations were contacted for, the length of time since they were last contacted, and the type of information they were contacted for in the past −

does not in general have a negative effect on survey response. Even in cases where negative effects were found, these were often small, leading the authors to formulate (in their title) an implicit hypothesis: "If we bother them more, are [the businesses] less cooperative?," implying that in fact the opposite might be true.

With the similar goal, Davis and Pihama (2009) carried out an analysis of establishment nonresponse on a mandatory annual survey using a number of explanatory variables. Among them was a measure of total survey burden in the preceding year. They found a statistically significant positive relationship between the total survey burden − expressed as a function of the total time taken in the year per 10 employees − and the probability of nonresponse in the survey, however they note that this effect was relatively small in magnitude in comparison to some other of the factors in the study (size of business, key provider status, stress the business might have due to expansion or contraction).

The studies above, which are the few we are aware of, were by the amount of available data restricted to a single survey or a small subset of all the data providers. In addressing these same issues, the current study utilises data from a register of participation of businesses in official statistics surveys kept at Statistics Sweden since 2009, which gives the possibility to analyse the phenomenon from the perspective of the total burden that a national statistical institute imposes on businesses for the purpose of data collection.

Our analysis is characterised by a presumed utility of observing two levels of business responding: on one level is the business's participation (response or nonresponse) in a specific instance of a specific survey: for instance, the April 2009 instance of Intrastat. However, all the survey instances that a business has received requests to participate in throughout a period of time lead to an aggregated level, the business level of response behaviour.

2. Data and Methodology

The register is officially known as the register of data provision (RDP). It keeps information on all direct data collections carried out by Statistics Sweden towards businesses, be it on appropriation or on commission from other producers of official statistics. Included is the information on whether a business has responded or not to any particular instance of data collection (conditional upon selection). Also available are properties of the selected businesses, such as their size or the industry they are active in, and to properties of the surveys, such as their periodicity, and whether they are mandatory or not. (In the current analysis, we use only 2009 data, as these data were the only available when the analysis was initiated.)

2.1 RDP content, filtering and categorisation

Data collection for the RDP is an automated process that "crawls" databases containing data on businesses sampled for the purpose of statistical production at Statistics Sweden. The databases contain, along with identifier details, information on whether a specific

business, for any specific instance of data collection it was selected for, has responded (provided data) or not. The RDP process collects data from the databases and from the metadata pertaining to the statistical programs in question. An additional part of the data is imported from other databases, prominently the Business Register (BR).

Data for 2009 consisted of 1,156,748 rows, each about a specific reporting unit's participation in a survey instance. Among the available variables were: legal unit ID, local kind of activity unit (LKAU) ID, business size (in number of employees), main industry in which the business is active (as a NACE code), survey ID, survey periodicity, whether the survey belongs to the official statistics of Sweden or not, whether responding to the survey is mandated by a law or regulation, date of sending the participation request to the business, date the response is due, average time it takes to provide data for the survey (estimated by the survey manager), and whether the response has been received.

Filtering the data, we removed entries which for different reasons did not have data on the key variables above, resulting in 1,055,698 entries. Further filtering restricted the data to the 14 largest surveys (among the 75 that existed prior to that): largest in terms of the number of entries in the RDP data set (accounting for more than 80% of the survey volume). This was done in order to have a reasonably low number of different surveys.

Thus, the data set that makes the corpus for the current study consisted of 845,812 entries pertaining to 168,151 businesses (defined as unique legal units) that were sampled for surveys by Statistics Sweden in 2009 for the 14 largest surveys. This number of businesses makes about one fifth of the total number of businesses registered in Sweden in 2008. There were no voluntary surveys among the 14 largest surveys: voluntary surveys were few (12 out of 75) and comparatively small in sample sizes.

Table 1: Codes used for the variable Industry.

Code	Provisional name	NACE-2 categories
1	Production	1 – 43
2	Trade	45 - 47, 68
3	Services (other than the Codes 4 - 6)	49 - 56, 94 – 96
4	Information and finance	58 – 66
5	Professional and administrative activities	69 - 82
6	Public welfare, etc	84 - 93, 97 – 99

We re-categorised two of the variables from the BR. The data on companies' size were re-coded into four categories of size: up to 4 employees, 5-19 employees, 20-199 employees, and 200 or more employees. The data on the main industry that the company was active in (originally being on a NACE-5 level) were re-coded into six coarse groups, presented in Table 1. Both re-categorisations were done in order to reduce the occurrence of empty or sparse cross-classified cells and to increase interpretability of the analyses.

2.2 Methods

To any particular survey instance, the sampled businesses respond or not, providing grounds for analysing survey instance response as a function of the businesses' properties

(like the size and industry that they belong to), survey properties (like periodicity and the average time that it takes to participate in its survey instance), and interactional properties (the number of surveys and survey instances the business is sampled in within the reference year, or a total response burden imposed during a preceding period). Business size class can additionally be seen as a proxy for survey effort. This analysis needs to be executed as hierarchical, as the businesses' properties are repeating their influence in the different survey instances that the business is invited to participate in, as well as a particular surveys' properties are repeating their influence in different survey instances. Thus, to account for the clustering of observations in the RDP into both businesses and surveys, we use a cross-classified multi-level model:

$$Pr(y_i = 1) = logit^{-1}(\boldsymbol{\beta X_i}),\text{ for } i \text{ indexing observations, with } \boldsymbol{\beta} \text{ including}$$
$$\boldsymbol{\beta}_j^{Surv} \sim N(\gamma^{Surv} \boldsymbol{u}^{Surv}, \sigma^2_{Surv}),\text{ for } j \text{ indexing surveys, and}$$
$$\boldsymbol{\beta}_k^{Bus} \sim N(\gamma^{Bus} \boldsymbol{u}^{Bus}, \sigma^2_{Bus}),\text{ for } k \text{ indexing businesses.}$$

Further, we analysed some of the data in the light of testing a null hypothesis that there is an independence between responses to survey instances by a specific business. To put it simply, whether the null hypothesis holds that, for every survey instance (of a specific survey) that the business is requested to participate in, the business decides whether to participate or not based on "tossing a coin" (but with probability of response not $P=0.5$ but that of the average response rate for the specific survey).

3. Results

3.1 Descriptive summaries

Basic demographics of the businesses in the 2009 version of RDP are, on the two levels of analysis –data collection instances and businesses – presented in Tables 2 and 3. The tables confirm the well known imbalance in data provision: while big businesses comprised only about 0.8% of the 168 thousand sampled businesses (Table 3), they were requested to provide about one fifth (20.7%) of the volume of data collection (Table 2), expressed as the number of survey instances. Correspondingly, small businesses comprised almost 70% of the sampled businesses but were together requested to provide just above one quarter (27.4%) of the data collection volume.

Table 2: Distribution of survey instances over NACE groups and business size.

Size: NACE group	1	2	3	4	Row counts	Row % to Total
Production	4.3	5.1	13.3	10.2	278874	33.0
Trade	10.0	8.5	7.6	3.1	247073	29.2
Other services	4.6	3.1	3.5	2.0	111578	13.2
Info. & fin.	1.6	0.9	1.7	1.7	49497	5.9
Professional	4.6	2.3	3.0	2.6	105871	12.5
Welfare	2.1	1.2	1.9	1.1	52919	6.3
Col counts	231419	178256	260792	175345	845812	
Col % to Total	27.4	21.1	30.8	20.7		100.0

Table 3: Distribution of sampled businesses over NACE groups and business size.

Size: NACE group	1	2	3	4	Row counts	Row % to Total
Production	13.1	5.3	3.0	0.3	36727	21.8
Trade	17.6	5.9	1.9	0.1	42959	25.5
Other services	11.2	3.6	1.4	0.1	27364	16.3
Info. & fin.	5.0	1.1	0.7	0.1	11577	6.9
Professional	15.4	3.0	1.2	0.1	33111	19.7
Welfare	7.3	1.5	0.9	0.1	16413	9.8
Col counts	117181	34303	15282	1385	168151	
Col % to Total	69.7	20.4	9.1	0.8		100.0

3.1.1 Response behaviour

On the survey instance level, surveys largest in terms of the volume of data collection instances are displayed in Table 4. Among them, Intrastat is largest with about 20 percent of the volume, as well as some data collections that provide statistics on short-term indicators of the economy (employment, prospects and turnover).

Table 4: Surveys included in the analysis and their response rates.

Survey name	Response rate	Row counts	Row %
Intrastat	92.1	173,457	20.5
Short-term employment	89.3	102,830	12.2
Short-term prospects on job openings	89.2	94,826	11.2
Turnover, month	77.0	82,119	9.7
Short-term prospects, salaries in private sector	78.7	73,874	8.7
Start-up businesses	72.4	68,583	8.1
Price indices in production and import	94.5	57,814	6.8
Salaries, longitudinal study	89.4	41,441	4.9
Occupation structure in Sweden	85.7	30,046	3.6
Lodgings in hotels, hostels and holiday villages	65.9	29,505	3.5
Short-term prospects in industry, month	83.9	26,992	3.2
Turnover, quarter	80.2	23,431	2.8
Investments in industry	82.0	20,536	2.4
Structural business survey	82.1	20,358	2.4
Col %	85.0	845,812	100.0

Response rates between surveys vary from 94.5% for a survey of price indices in production and imports to 65.9% for a survey on lodging statistics. While of interest for improving the data collection instruments or for further work on improving motivation of the businesses for responding, analysing this was beyond the scope of the present paper.

Survey instance response rates (Table 5) vary with business size from 73.2% for small businesses monotonically to 92.9% for biggest businesses. This likely includes an effect of the somewhat varying survey effort to obtaining responses, which is likely related to the sampled businesses' inclusion probabilities. Response rates vary also over NACE categories on the level of data collection instances, from the lowest of 76% in services to 89% in production (displayed in Table 6).

Table 5: Survey instance response rates per business size.

Size group	Response rate	Row counts	Row %
1	73.2	231419	27.4
2	85.4	178256	21.1
3	90.0	260792	30.8
4	92.9	175345	20.7
Col %	85.0	845812	100.0

Table 6: Survey instance response rates per NACE category.

NACE group	Response rate	Row counts	Row %
Production	88.7	278874	33.0
Trade	85.2	247073	29.2
Other services	76.2	111578	13.2
Info. & fin.	85.3	49497	5.9
Professional	85.0	105871	12.5
Welfare	82.7	52919	6.3
Col %	85.0	845812	100.0

On the business level, of all the sampled businesses, 118,395 businesses (70.4%) always responded, 18,812 businesses (11.2%) sometimes responded[1] and 30,944 businesses (18.4%) never responded. Those that never responded were in the vast majority small firms: either in size class 1 (0-4 employees) (86%) or in size class 2 (5-19 employees) (12%). Only two large businesses (200 or more employees) were among the never-responders. However, the similar holds for those businesses that always responded: these were also mostly businesses in size classes 1 (82%) and 2 (16%).

Table 7: Business level, number and percent of surveys requested to participate in, with a breakdown by participation pattern.

# of surveys	Never-responders		Sometimes-responders		Always-responders	
	Counts	%	Counts	%	Counts	%
1	28,856	93.3	4488	23.9	97237	82.1
2-3	2,015	6.5	9355	49.7	18959	16.0
4-6	72	0.2	3725	19.8	2126	1.8
7-10	1	0.0	1244	6.6	73	0.1
Sum	30,944	100	18812	100	118395	100

Most of the never-responding businesses have been asked to participate (Table 7) in only one survey (93% of the never-responders); only very seldom (0.2%) in four or more surveys. Displayed in Table 8, most of the surveys that the never-responding businesses were asked to participate in had a single instance (that is, they were either annual surveys or intermittent surveys): 86% of the never-responders were asked to participate only in such surveys. For less than 1% (more precisely 0.6%) of the never-responding businesses

---

[1] Sometimes-responding is defined as having a response rate (number of survey cycles responded to/number of survey cycles invited to participate into) between 0 and 1, excluding the endpoints. With this definition, only businesses receiving invitation to two or more survey cycles could be in this category.

would participating in more than 12 survey instances have been required. Comparative data for expected length of survey participation are presented in Table 9.

Table 8: Business level, number and percent of surveys instances requested to participate in, with a breakdown by participation pattern.

	Never-responders		Sometimes-responders		Always-responders	
# of surveys instances	Counts	%	Counts	%	Counts	%
1	26,674	86.2	0	0.0	87719	74.1
2-12	4,090	13.2	10824	57.5	25225	21.3
13-120	180	0.6	7504	39.9	5448	4.6
121-499	0	0.0	440	2.3	3	0.0
500 and more	0	0.0	44	0.2	0	0.0
Sum	30,944	100	18812	99.9	118395	100

Table 9: Business level, time category and percent of surveys requested to participate in, with a breakdown by participation pattern.

	Never-responders		Sometimes-responders		Always-responders	
Total time (in minutes)	Counts	%	Counts	%	Counts	%
Up to 15	22,652	73.2	0	0.0	68877	58.2
16-60	3,911	12.6	1936	10.3	20088	17.0
61-120	2,743	8.9	2046	10.9	13305	11.2
121-600	1,019	3.3	5978	31.8	9359	7.9
601-2,400 (one week)	612	2.0	6666	35.4	5406	4.6
2,401-10,000 (one month)	7	0.0	2112	11.2	1360	1.1
10,001 and more	0	0.0	74	0.4	0	0.0
Sum	30,944	100	18812	100	118395	100

3.2 Analytical modelling

3.2.1 A hierarchical model for predicting response in a survey instance

Using the general model specified in Section 2.2, a number of models were tried, containing predictors on the levels of businesses (size, industry, number of units to respond for), surveys (length of time it takes to fulfil the request, periodicity (monthly or not)), and interaction between business and survey organisation (i.e. total response burden imposed on the business, in terms of the total number of surveys, survey instances and total time needed to fulfil these).

Further, in order to simplify interpretation of the total imposed response burden (which per definition includes the current survey request), the annual data were separated into two halves, the total burden (in terms of time) was calculated for the first half of the year and used as a predictor in a model where the response was participation (yes=1, or no=0) on survey instances sent out during the second half of the year. (To simplify computational considerations, we have excluded businesses that contribute with only one observation; there were 344,988 observations in the resulting data set.)

Based on likelihood ratio statistics and the associated p-values of the models (Pinheiro and Bates 2000), the model finally chosen was the one displayed below.

```
              Estimate Std. Error  z value  Pr(>|z|)
(Intercept)   1.152998   0.287033    4.017  5.90e-05 ***
Size2         0.867569   0.051239   16.932   < 2e-16 ***
Size3         1.071128   0.067377   15.898   < 2e-16 ***
Size4        -0.175042   0.767280   -0.228  0.819543
Burd2        -0.585823   0.076381   -7.670  1.72e-14 ***
Burd3         0.403873   0.078451    5.148  2.63e-07 ***
Burd4         0.675385   0.168957    3.997  6.41e-05 ***
Burd5        -4.762237   0.233613  -20.385   < 2e-16 ***
Ind2          0.206942   0.041443    4.993  5.93e-07 ***
Ind3         -0.474897   0.050121   -9.475   < 2e-16 ***
Ind4          0.029930   0.065932    0.454  0.649859
Ind5          0.198842   0.051388    3.869  0.000109 ***
Ind6         -0.062065   0.062615   -0.991  0.321583
Size2:Burd2   0.112017   0.094706    1.183  0.236891
Size3:Burd2   0.480057   0.103148    4.654  3.25e-06 ***
Size4:Burd2   1.805936   0.802478    2.250  0.024420 *
Size2:Burd3  -0.004698   0.100968   -0.047  0.962890
Size3:Burd3  -0.118630   0.111720   -1.062  0.288304
Size4:Burd3   0.754264   0.779930    0.967  0.333498
Size2:Burd4   0.229574   0.206157    1.114  0.265456
Size3:Burd4  -0.018107   0.184612   -0.098  0.921869
Size4:Burd4   0.735123   0.795896    0.924  0.355673
Size2:Burd5   3.630175   0.293659   12.362   < 2e-16 ***
Size3:Burd5   4.744150   0.231347   20.507   < 2e-16 ***
Size4:Burd5   6.086467   0.798387    7.623  2.47e-14 ***
```

The model includes the business level variables *Size* and *Industry*, the imposed response burden variable *Burden*, and an interaction between *Size* and *Burden* (in order to control for the possibly differential survey effort, which is commonly related to size/inclusion probability). Do note that no survey-level variables were deemed needed.

Table 10 summarises the interactions between the Size and Burden covariates. Each cell is the sum of four estimated effects: the intercept, the *Size* class effect, the *Burden* class effect, and an interaction of the latter two. (The first row and column correspond to the reference categories for *Size* (class 1) and *Burden* (class 1) respectively, thus their effects and interactions are zero; in particular, cell (1,1) contains only the intercept. This table does not include the *Industry* variable, whose coefficients are not large and do not interact in this model with the other explanatory variables.)

Table 10: A response propensity surface based on the point and variance estimates of the hierarchical model, with standard errors of the estimates in parentheses.

	Burd1	Burd2	Burd3	Burd4	Burd5
Size1	1.153	0.567	1.557	1.828	-3.609
	(0.287)	(0.292)	(0.293)	(0.323)	(0.366)
Size2	2.021	1.547	2.420	2.926	0.889
	(0.288)	(0.289)	(0.291)	(0.308)	(0.385)
Size3	2.224	2.118	2.509	2.881	2.206
	(0.291)	(0.288)	(0.291)	(0.290)	(0.311)
Size4	0.978	2.198	2.136	2.388	2.302
	(0.818)	(0.363)	(0.310)	(0.314)	(0.298)

In any size class, extreme response burden (here, the uppermost 20% of the burden distribution) leads to a reduction in the ability to provide data − this effect stronger in small businesses than in large, indicating an interaction: large businesses may be hypothesised to be less fragile and are thus better able to withstand extreme response burden; or, as indicated before, they might be exposed to higher levels of survey effort. But, over all the size classes, "the more we burden them" does seem to have a negative effect in terms of response rates.

3.2.2 Testing independence of responding on survey instances

Investigating independence between participating in survey instances by a business (see Section 2.2), we used data from two specific monthly surveys: "Turnover statistics" and "Short-term prospects in industry". For each survey, we looked at only newly-selected businesses, noting the number of times they have responded in their first seven and eight, respectively, consecutive survey instances.

If decision on participation is independent within each business and for a specific survey, then the number of times that the business participates follows a binomial distribution. We also assumed that there are a number of deterministic non-respondents who never respond. All the other units are seen as stochastic responders who respond to the survey rounds in a stochastic way. (Those deterministic non-respondents have to be estimated and treated separately, as we elaborated in Chapter 2 of Giesen and Bavdaz (2013).)

Figure 1 displays, on the left, a comparison between the observed (in the upper row) and the expected under independence (in the lower row) distributions for "Turnover statistics"; and on the right the same comparison for "Short-term prospects". It is obvious even by simple visual comparison that the upper and lower distributions in any of the two pairs strongly differ. Therefore, participation is not independent of past participation. From the current analysis, it is not possible to deduce whether the dependence is positive (previous response raises future response e.g. due to learning and adaptation) or negative (previous response lowers future response e.g. as result of accumulated burden).
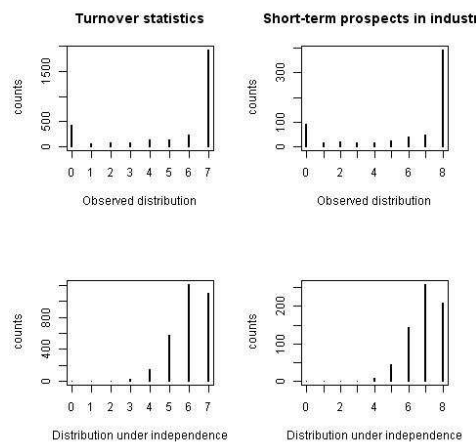

Figure 1 - Comparisons for the independence tests.

4. Summary

The descriptive presentations based on the RDP indicate usefulness of a register of data provision in order to gain a holistic picture of businesses' involvement in providing data for official statistics. A small proportion of the businesses were required to provide a considerable amount of human resources to comply with the requests, up to ten person-months in one case. While it was not the goal of the current analysis to investigate reasons for such occurrences, one may note that − in order to improve precision of estimates − sampling of businesses is sometimes carried out in a so-called positively coordinated way, resulting in businesses remaining in samples for extended periods. The RDP or similar register may provide a way of identifying and rectifying large variations in response burden businesses are exposed to.

Using a multi-level analytical model, we gathered further evidence that too high levels of response burden on a business may lead to the business's reduced propensity to participate when requested to subsequently provide new data in survey instances, and possibly to other deteriorations in data quality.

Businesses' response behaviour between the survey instances is not independent: in particular, businesses that are nonrespondents in the first instance of a survey do run a risk of remaining nonrespondents and thus are in need of particular attention to be paid to them by the data collecting organisation.

References

Davis, W.R., and N. Pihama (2009). Survey Response as Organisational Behaviour: An Analysis of the Annual Enterprise Survey, 2003–2007. Paper presented at New Zealand Association of Economists conference, Wellington, New Zealand, 1–3 July 2009. Available online: http://www.stats.govt.nz/surveys_and_methods/methods/research-papers/nzae/nzae-2009/~/media/Statistics/surveys-and-methods/methods/research-papers/NZAE/2009/nzae09-survey-response-org-behaviour.pdf (accessed 2013-01-29)

Giesen, D., and M. Bavdaz (Eds.) (2013). BLUE-ETS Deliverable 8.1: "Comparative report on integration of case study results related to reduction of response burden and motivation of business". To be made available at www.blue-ets.eu.

McCarthy, J.S., Beckler, D.G., and Qualey, S.M. (2006). "An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?" Journal of Official Statistics, 22(1), 97–112.

Pinheiro, J.C., and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS. Springer, New York.