**Could "fresh" register data improve answers in surveys?**

Øyvind Bruer-Skarsbø, Statistics Norway, obr@ssb.no

Kristin Halland, Statistics Norway, knh@ssb.no

**Abstract**

*Survey data from establishments are notoriously flawed with logical errors that are easy to identify after the collection of data is complete. These errors typically stem from misunderstandings about the meaning of the questions, or about which unit to report for. Logical errors have previously been identified by using register data that arrive after the survey has taken place. The handling of these errors is time-consuming, costly and detrimental to the data. Statistics Norway (SSB) has recently implemented out a way to handle this problem in the quarterly survey on self-certified sick leave, made possible by the ever “fresher” availability of register data. By re-using data about the number of women and men employed in the establishment, SSB tries to make clearer which unit the respondents are reporting for and create a dynamic and coherent set of controls that makes it easier for the enterprise to understand the questions. The result being that many of the controls can be moved out of SSB and into the hands of those best qualified to answer the questions. Hopefully, the results being that the burden on establishments, and costs for SSB, is reduced.*

*In a world where the amount of available data is greater, and the frequency at which it is available to statistical offices is higher, it should be used to improve the quality of survey data and to reduce the burden on establishments. Norway has found a way to utilize its increasingly “fresh” register data on employees to do this when collecting data that is not (currently) available through a register. This paper will present the background, implementation and results of this work. The results should be useful to other countries that have the same access to data, or that may have in the future.*

**1. Introduction**

Despite the growing availability of register data in Norway, there are still surveys that collect data on variables that are "register-friendly". One such example is the survey on self-certified sick leave at Statistics Norway (SSB). Businesses are required by law to keep track of self-certified sick leave amongst their employees, and the information is to a large degree digitally available in their payroll systems. But for various reasons a register has been difficult to get into place, and a survey is still necessary. Since this way of collecting data is costly for both businesses and producers of official statistics, it's important to explore all channels that can reduce these costs. One such approach, re-using information about the establishments to improve answering, has been implemented in the survey on self-certified sick leave at SSB.

This paper discusses the background, implementation and preliminary results of the method implemented in the survey on self-certified sick leave at SSB. Starting from 2015, Norwegian enterprises are obliged to send a monthly report with detailed information about their employees to be used by the Norwegian Tax Administration, Norwegian Labour and Welfare Administration (NAV) and SSB. SSB uses this information to create dynamic controls directly in the questionnaire, and to inform the enterprise about which establishment it should report for.

**2. Background**

Since 2001 SSB has published quarterly statistics on sickness absence. The data stems from two sources:

1. Register of doctor-certified sickness absence
2. Survey of self-certified sickness absence

The first is a register based on data sent directly from the doctors who certify sickness absence, and is administered by NAV. The second, the focus of this paper, is a survey of about 10 000 establishments in Norway. In this chapter, the peculiarities of this survey will be presented. This should create a background for understanding how improvements can be made in process of getting hold of data.

*2.1. The questionnaire*

The statistics on sickness absence publishes two main quantities in relation to sickness absence; the number of lost man days and the number of individuals who has been sick, either as percentage of total sum of man-days, or as an absolute number. In the case of self-certified sickness absence, SSB only publishes the number of lost man-days as a percentage of the total number of man-days.

In the questionnaire, the respondents are asked to answer the following seven questions about the sick leave in their establishment:

1. Can the employees be absent more than 3 calendar days in one self-certified sick leave? Yes/No
2. Please state the number of calendar days a person can be absent in one self-certified sick leave.
3. Has there been any self-certified sick leave in the period? Yes/No
4. State the number of men and women who have had at least one self-certified sick leave in the period.
5. How many self-certified sick leaves where started by these <X> individuals in the period, for each sex?
6. How many working days in total, for each sex, were lost due to self-certified sick leave in the period?
7. How many man-days, for each sex, did these <Y> lost working days amount to?

Questions 1 and 3 are path-dependent, meaning that the answer given here will decide which questions are next. Question 1-2 is related to the Norwegian IA Agreement[[1]](#footnote-1) and serves as a control variable later in the questionnaire. The main question for publishing the statistic is question 7, and the others are designed to make the respondent understand the meaning of this question.

As is evident, there is a sort of hierarchy in the ordering of questions 4-7. They are set up so it's easy to address logical errors in real-time while the respondents are answering. This will be an important point later in using register-data for improving data quality.

*2.2 Data collection*

Data is collected through an internet portal for digital dialogue between enterprises, private individuals and public agencies, called Altinn. Since many of the public agencies use this portal for sending out and gathering information, the enterprises are very well acquainted with using it. SSB also uses this portal for most of its surveys, and businesses transmit their monthly report on employees to a-ordningen in this portal. The data from a-ordningen is central to this paper and will be discussed in more detail in the next chapter.

Around 90 percent of the sample answers the survey manually (a physical person answers the questionnaire), and around 10 percent answer automatically. The latter meaning that the payroll systems generates an automatic report that is sent every quarter to Altinn. These two ways of submitting answers obviously has implications for the types of errors one expects. The focus in this paper is on the 90 percent that submits the answers manually.

*2.3 Challenges for the data*

A fundamental problem for all establishment-based surveys at SSB is that this unit is often not well defined in the enterprises themselves. Answering the survey is a legal obligation for the enterprise, the legal unit. But for statistical purposes, we are often interested in the different industries and/or locations within an enterprise. Thus, we want the enterprise to answer for specific parts of the enterprise, parts that they might not recognize. In other words, when answering the questionnaire, the respondent is sometimes confused about which employees' it should report for. This is of course a serious problem and can lead to measurement errors in data.

In the survey on self-certified sick leave SSB asks the enterprises (legal unit) to answer for one or more of its establishments (sub-unit). These sub-units are mainly constructed to cover the needs of public agencies, i.e. for statistical purposes. The person answering the questionnaire might not recognize this structure, and only know the internal (or operational) structure of the enterprise. This generates confusion about which of the employees they should report sick-leave for. This confusion can lead to large errors that SSB has had to handle after the data collection process has ended.

It might seem puzzling that establishments have problems knowing how many employees they have when answering the survey, given that they report this to a-ordningen each month. One possible explanation could be that reporting to a-ordningen mostly happens automatically via payroll systems. The person responding to the survey might not be aware of this. In fact, we often experience that respondents don’t recognize the establishment-identifier. For large enterprises, this identifier is crucial for creating reports about the sick leaves of the employees in different establishments.

**3. A-ordningen**

Since 2015 the Norwegian Tax administration, NAV and SSB have received monthly data from all Norwegian establishments with information on jobs, renumeration in cash and kind, and taxes. This coordinated way of sending information about employees from businesses to the public agencies, is called a-ordningen. It forms the basis of SSB`s detailed register-data about all Norwegian employees (hereby this register-data is also referred to as a-ordningen). This kind of data is nothing new in itself, but the precision and the speed at which it is available for (re-)use creates new opportunities.

*3.1 Timeline of the data*

Since the survey on self-certified sick leave is quarterly, and it makes the most sense to use information on employees from the same period for our purposes, the timeline in the production of a-ordningen is crucial.

The raw data for month t arrives at SSB around the 10th of month t+1. After being subjected to some various processing, a first version is available around 5 to 10 days after the arrival of the raw data. This is often referred to as version 1 of month t. If the establishments wish to make corrections to data sent in for month t, it can do so in the reporting for period t+1. Then a version 2 of month t is created and is available for use around the 15th of month t+2. Since version 2 is more precise than version 1 (especially with respect to the number of employees), the survey on self-certified sick leave use the number of employees from this version in the questionnaire. The survey is sent to the respondents a few days after the end of the quarter, around two weeks after the version 2 for the first month of the quarter is available.

The number of employees from a-meldingen is not only used for controls in the questionnaire. Its also used to estimate the population levels of self-certified sick leave, and for stratification when drawing the sample. The latter has some relevance for this paper. Every year part of the sample is replaced by new establishments. To reduce the burden on small enterprises, we do not include establishments with less than 5 employees. After the enrolment of new establishments, SSB usually receives a lot of applications from the new units wanting to be exempted from participating. A lot of these applications are turned down with the argument that they do not satisfy the 5-employee-cutoff. This leads to discussion about how many employees the establishment has. This has proved to be useful knowledge when designing the questionnaire.

Register-data from a-ordningen is valuable for many reasons, one of which is that it can be used to improve the quality of official statistics. With respect to surveys, it has the potential to both reduce burden of the respondents and improve the quality of the answers. It has already been implemented in the data collection process in several statistics at SSB. In the case of self-certified sick leave, it was implemented in the survey for the first time in 4th quarter of 2018, confronting the respondent with the number of employed men and women in the first month of the quarter under question.

*3.1 Relevance for self-certified sick leave*

A-ordningen contains precise information about the number of employees in all Norwegian establishments. Since the number of employees is positively correlated with the number of employees being sick in a given quarter, it has been used in the post-data-collection editing for a long time (before a-meldingen the old register was used), i.e. to identify outliers and logical errors. It is also used as an estimator for inference about the population level of self-certified sick leave. But recently the rate at which this data is available for re-use by SSB, has increased. That enables us to confront respondents with the number employees they themselves reported to be working there in the period we are asking for. This is a vast improvement compared to how this information has traditionally been used, where data from previous periods where used.

Starting from the 4th quarter of 2018, SSB stated the number of employed men and women the establishment had reported in for the first month of the quarter. This information is displayed after question 3 in the questionnaire (see chapter 2.1). Doing this has two great advantages:

1. It clarifies which sub-unit/establishment the respondent should report for (as discussed in chapter 2.3).
2. It enables SSB to build a coherent set of controls around precise information given by the establishment itself.

The first point addresses the concern raised in chapter 2.3. Being clear about which group of employees we want information about, is crucial for the data quality. The second point speaks to the fact that this information enables SSB to move controls out of SSB and into questionnaire. Both have the potential to reduce costs for both SSB and the respondent.

*3.2 Challenges with using a-ordningen*

Although a-ordningen is a great source for precise data about employees, it also has some issues that has had to be dealt with when using it in the manner described above. Firstly, for most establishments the transmission of the data is done automatically by the payroll system of the enterprise. This has a couple of consequences:

* SSB receives far more jobs[[2]](#footnote-2) among employees than is likely. A lot of these might be "old" jobs which are not removed from the payroll system, or they may be jobs where the employee has not been payed that month. For that reason, SSB must "wash away" a lot of the jobs. This might affect how clarifying these numbers are for the respondent in the survey. If the respondent knows what was reported to a-meldingen, they might not recognize the number of employees.
* Since reporting to a-meldingen often happens automatically, the person answering the questionnaire might not relate to the reported number of employees.
* There might be large changes in the number of employees between the first and last month of the quarter.

The way SSB has dealt with these challenges, is to carefully formulate the text that explains how the number of employees is derived. The resulting text, in our opinion, should clarify all these types of misunderstandings. Judging by the number of calls SSB has received after the deployment of the new questionnaire, it has not been a big problem. In fact, the impression has been that it has had the opposite effect. If the establishment finds the stated number of employees unlikely, they might look further into what their system is reporting to a-meldingen and correct any wrong-reporting by their system. This is typically the case after mergers and acquisitions.

**4. Implementation**

Starting from the 4th quarter of 2018, SSB confronted the respondents in the survey on self-certified sick leave with how many employed men and women they themselves had reported for the first month of the quarter. Also, several controls were introduced that were based on the number of employees. These controls were for the most part not new. The new thing was that the controls got moved out of SSB and into the questionnaire. In general, SSB aim to control and validate as much as possible on the respondent side, i.e. before a questionnaire is submitted. But doing this is not straight forward, and several considerations and choices had to be made. The most important of which are discussed below.

*4.1 Confusion or clarity*

The most basic consideration was whether this change would bring more confusion than clarity for the respondent. The reason for this has been discussed earlier and has to do with the definition of establishment and how the number of employees are reported automatically via payroll-systems. If it were to confuse more than clarify, adding controls would only make things worse. Given SSBs experience in dealing with applicants wanting exemption from the survey, the decision was made that it would clarify more than it would confuse. The key insight being that there needed to be no confusion on which jobs were included in the figures respondents were confronted with. In SSBs experience, people often tend to think of employees as only consisting of full-time workers. But the figures presented to them are (almost) all workers who have received wages in the period.

*4.2 Setting up the controls*

Since the number of employees are positively correlated with sick-days, it’s a useful variable for controlling the answers. Several soft controls were set up in the new questionnaire that were based on the number of employees. The controls are “soft” in the sense that activating one will not stop you from submitting the questionnaire as with “hard” controls. The reason that this type of control was implemented is that we only use information from the first month of the quarter. In some cases, the number of employees vary greatly from one month to another.

After question 3 in the questionnaire, the number of employed men and women they reported to a-meldingen for the first month of the quarter, is presented to the respondent, followed by a short explanation of which employees are included in the figures. The respondent can also click on a link to get more detailed information.

Question 4 is about how many men and women that have been sick at least one time in the period. Since having more sick people than employees in the first month of the quarter is likely to be an error, this question has a control against the number of employees. If the respondent triggers the control for men, the following text appears in the questionnaire:

“You have reported a high number of men with self-certified sick leave in the period. Based on information from a-meldingen you had <X> male employees in the first quarter of the month. Please check that the reported number of men with self-certified sick leave is correct.”

Question 5 asks about the number of self-certified sick leave cases that started in the period, with a control that builds upon the last control. Since one would expect that each person has at least one sick leave case, there is a control which is triggered if the number of sick leave cases are lower than the number sick individuals.

Question 6 ask about the number of working days lost due to self-certified sick leave. The reported amount triggers a control if the following criteria is met:

Lost working days > number of men (or women) \* Q2

Were Q2 refers to the answer given in question 2.

The last question concerns the number man days-lost due to sick leave. If the stated number is higher than what was answered in question 6, a control is triggered.

As is evident, there is a hierarchy of controls, starting with a control directly against the number of employees from a-meldingen. This is much like the post-data-collection editing that is done by SSB. But now we can confront the respondent directly and hopefully improve the quality of the answers. In the post-data-collection editing we also control for large changes in the number employees during the period.

*4.3 Satisficing*

A potential problem when introducing any control in a questionnaire is satisficing. If the respondents view an error message as an obstacle to be overcomed, they might be tempted to answer something that doesn’t trigger the controls (and not correct). This is the problem of satisficing. Since a hard control would probably lead to more of this than a soft control, SSB decided on the latter. Overall, the benefits of moving controls into the questionnaire outweighed the possible degradation in data quality from satisficing.

There is always a risk associated with using register data in questionnaires. In this survey we suggested that the benefits were greater than the risk.

**5. Results**

Judging whether the changes in the questionnaire has been successful or not, is not a straight forward matter. Ideally one would want to measure the behaviour of the respondents directly in the questionnaire. Unfortunately, the feature of tracking behaviour in the questionnaire was not available to us at the time of deployment. In the following we therefore focus on the data that has been submitted, comparing data before and after the change.

Since we barely have two periods to analyse (answers for the 1st quarter of 2019 is not completed), the data is somewhat limited. Still, in this chapter we try to judge how successful the changes have been by focusing on these measures:

1. How well does the first month of the quarter represent the number of employees for the whole period?
2. Has it reduced the number of (logical) errors?

The first point will give us an idea of how many establishments have large variations in the number of employees during the quarter, making it less correlated with sick leave and perhaps more confusing to relate to in the questionnaire. The second point is a check of how many times the respondents triggers the controls. Again, we only have the final version they’ve submitted and can't say anything about whether the controls caused the respondent to change their answer.

*5.1 Using the first month in the quarter*

When we decided to make this change in the questionnaire we had three options about which employee-figures to use:

1. Version 1 of the first month of the quarter
2. Version 2 of the first month of the quarter
3. Average of version 1 from the first and second month of the quarter.

As discussed previously, the decision was made that version 2 of the first month of the quarter should be used. The reason being that version 1 would probably include too many employees, and an average over two months with version 1 would have included too many employees and been harder to interpret/check for the respondent. Figure 1 shows how representative the first month is for the rest of the quarter. This is measured by absolute deviations in the number of employees between the first month and the average over the 3 months in a quarter.

**Figure 1. Absolute deviations in the number of employees between the first month and average over the quarter.**

**Source: Statistics Norway.**

From the graph we see that between 30 and 40 percent of the establishments has no change in the number of employees during the quarter. Around 50 percent has an increase or decrease between 1 and 5. As expected, the number of employees is not very volatile for the great majority of establishments. In the other end of the scale, between 3 and 6 percent of the sample experience quit large changes, from 21 and upwards. But the consequences of confronting the respondent with the number of employees for the first month of the quarter, if the number of employees is very volatile, should not be severe.[[3]](#footnote-3) It’s stated very clearly in the questionnaire that the figures are from the first month of the quarter. However, in our experience these establishments are either undergoing reorganisations, or experience large seasonal variations. The latter group likely consists of employees who might not qualify for self-certified sick leave (part-time workers, seasonal workers, etc.). This could potentially lead to these establishments contacting SSB to ask about these figures or apply for exemption, which can be costly for both parties. Judging by the first two quarters, this has not been the case.

*5.2 High probability errors*

To get a preliminary measure of whether the changes made have improved data quality, it seems natural to check the frequencies of answers which has a high probability of being wrong. Question 4 is the one which correlates the strongest with the number of employees, and it seems like a natural candidate for a quality check. Figure 2 shows the number of establishments which triggered a strong version of the control in question 4. It’s strong in the sense that these are only establishments which reported more sick men and women than the highest number of employees during any month of the quarter. This means that they are likely errors not caused by a volatility in the number of employees.

Looking at figure 2 we see that the level of these types of errors were relatively small before and after the change, with a small dip after the implementation (2019 for the 1st quarter and 2018 for the 4th quarter). The interpretation of these results is not a given.

**Figure 2. Number of establishments triggering a strong version of control in question 4.**

**Source: Statistics Norway.**

Hopefully the reduction in these types of error lead to less measurements error. In general, a fruitful way to evaluate new methods is to identify how it impacts the end results – i.e. the figures that get published. Unfortunately, considering the impact on estimation, seasonal adjustment and the published figures is beyond the scope of this paper. But any reduction in these extreme cases should be viewed as positive for data quality.

**6. Conclusion**

Two quarters after the implementation of a new method for improving quality in the survey on self-certified sick leave, its hard to make any hard claims about the effect of the change. Still, SSB finds the method important for two primary reasons. Firstly, since we are seeking answers about a unit which is not well defined for the respondent, we owe it to them to be as clear as possible about which unit(s) we want them to answer for. Using information given by the enterprise itself two months prior, which is directly connected to the unit they are asked to answer for, seems like a big improvement. On the one side it could reduce the burden for the respondent, on the other side it has the potential to improve quality of data. Secondly, the high quality of this register data lends itself well to moving controls out of SSB and into the questionnaire. This also has the potential to improve quality of data.

1. An agreement that aims to facilitate a high degree of employment in Norway. One of the means to do this has been to allow for a higher number of possible self-certified sick-leave days. [↑](#footnote-ref-1)
2. Jobs in this paper are defined as jobs among wage earners and does not include jobs among the self-employed. [↑](#footnote-ref-2)
3. To be on self-certified sick leave, an employee must have worked two months at the current employer. Thus, the newly employed will rarely be eligible for self-certified sick leave in a given quarter, if they were hired in that same quarter. [↑](#footnote-ref-3)