

On the relationship between response burden and data quality

A complementary study of response burden and data quality in official establishment surveys

Quest 2015 Workshop, Helsinki 28th-30th of April



Statistisk sentralbyrå
Statistics Norway

Kari-Anne Lund and Gustav Haraldsen
Department of Data Collection and Methods, Statistics Norway

Background

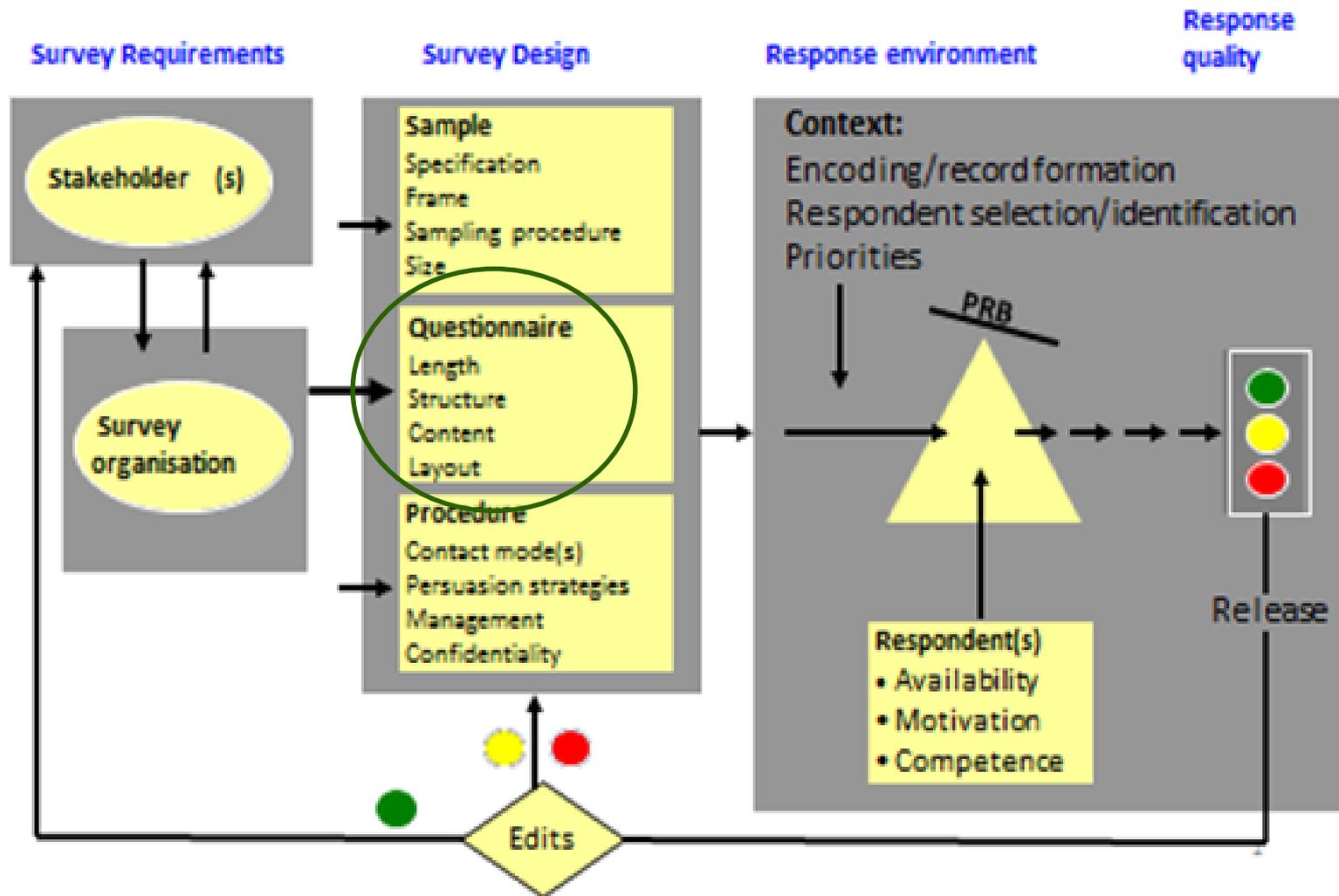
- Investigate and estimate the correlation between response burden and data quality
- Inconclusive results from previous studies
- Questions on response burden at the end of several official establishment surveys
- Results could affect where we put our efforts in developing the questionnaires



High response burden equals
poor data quality?



Origins and Effects of Perceived Response Burden in Establishment Surveys



Response burden questions

1. Did you find the questions in the questionnaire easy or difficult to respond to? (Easy, neither easy nor difficult, difficult)
2. Which questions were particularly difficult to respond to? (open text field)
3. Did you find working with the questionnaire easy or difficult? (Easy, neither easy nor difficult, difficult)
4. Which conditions contributed to make the questionnaire difficult?
(Tick of one or several conditions)

- Many questions → Length
 - Untidy layout made the questionnaire heavy to read → Layout
 - Difficulties finding out how the web questionnaire worked → Structure
 - Functionalities in the web questionnaire did not work as intended
 - Unclear concepts and explanations of concepts
 - Information that was not accessible before the reporting deadline
 - Difficult or time consuming calculations related to questions
 - Difficulties deciding on which response category or response was the correct one
 - Questions about information that do not correlate with the information we have in our systems
 - Information that I needed help to get access to or finding
 - Other reasons. Please specify:
- Content**
- (Information sources)**
- Had no problems with the questionnaire



Conceptual Model

| | <i>Case A</i> | <i>Case B</i> | <i>Case C</i> |
|---|---|--|--|
| <i>Survey</i> | Wage survey 2008 | Structural statistics 2010 | Resource depreciation 2013 |
| <i>Target group</i> | Establishment survey | Establishment survey | Establishment survey |
| <i>Mode, sample size</i> | Web self-reporting N=150 | Web self-reporting N=16 572 | Web self-reporting N=857 |
| <i>Response burden indicators</i> | Perceived response burden | Perceived response burden | Perceived response burden, amount of time spent reporting, accept of deadline, |
| <i>Data quality indicators</i> | Occurrences of automatic logical control violations | Number of manual edits, number of sent reminders | Missing, internal consistency, errors |



Overview of indicators

Response burden

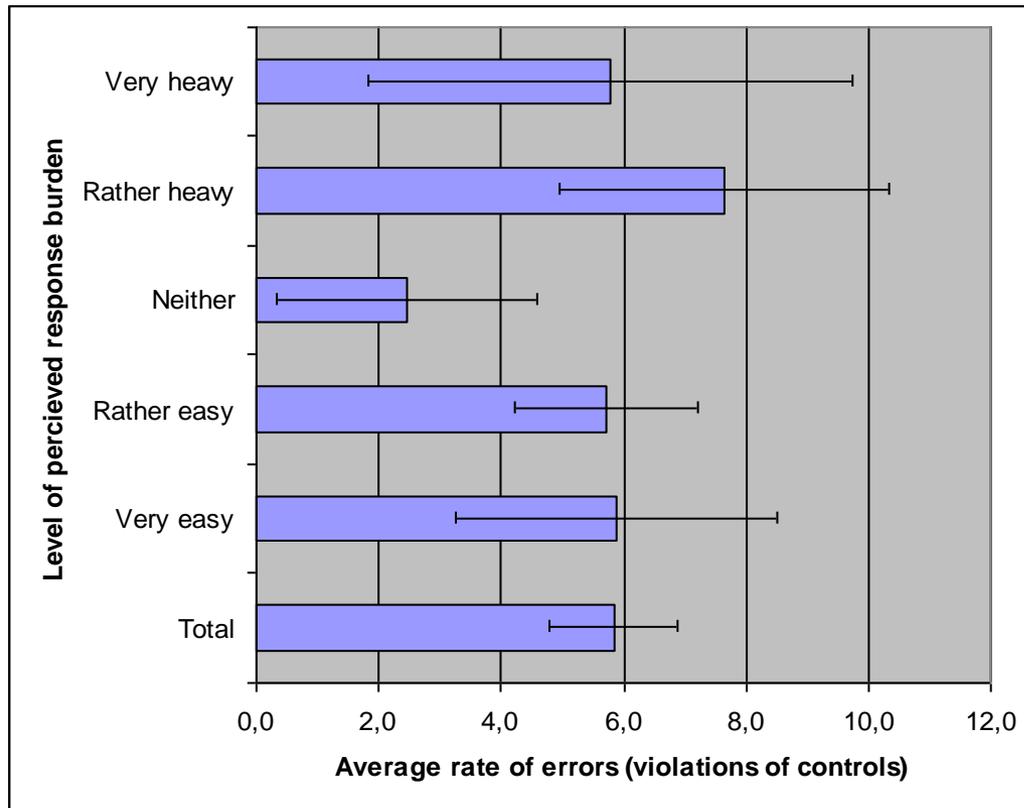
- *Perceived response burden*
- Amount of time spent reporting (C)
- Acceptance of deadline for reporting (C)

Data quality

- Missing values
- Internal consistency
- Obvious errors
- Manual edits
- Number of reminders sent
- Automatic control violations



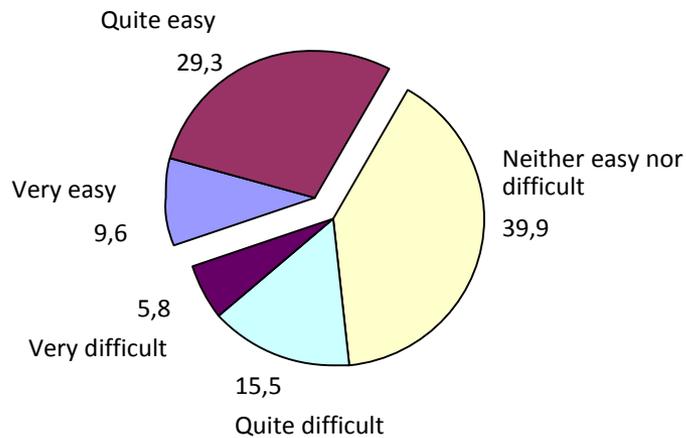
Case A: Occurrences of *automatic control violations* by perceived response burden (PRB)



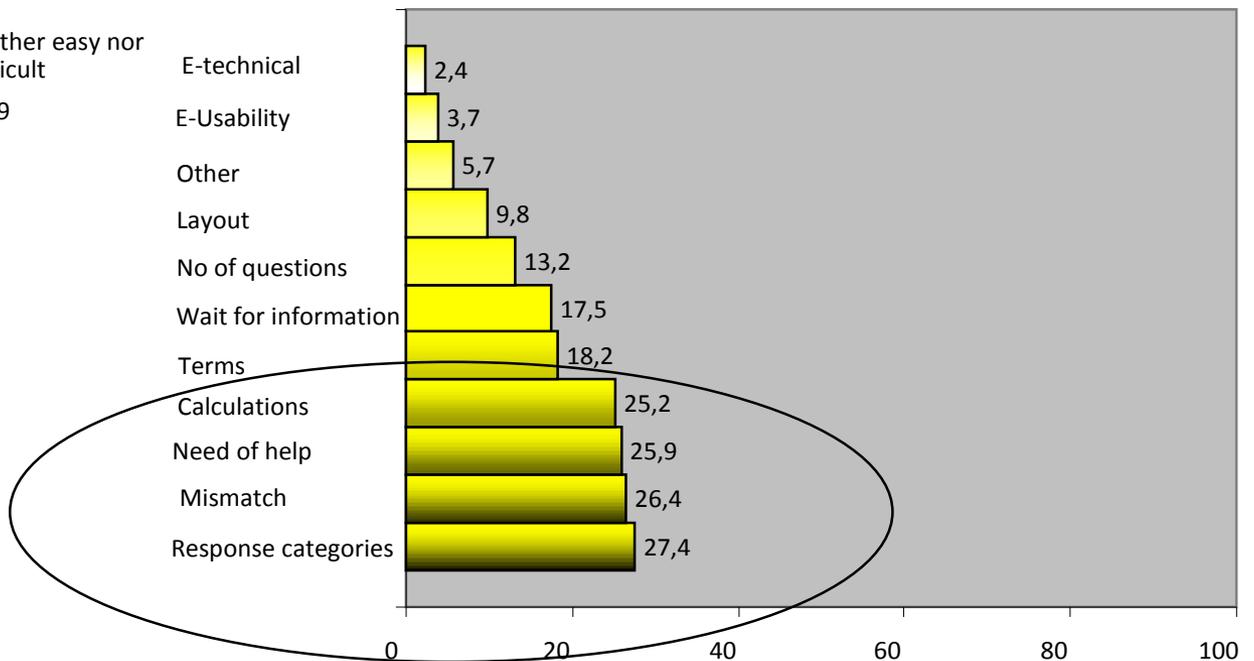
No significant connection between PRB and control violations

Case B. Perceived response burden and sources of response burden. Structural statistics 2010. Percent (n = 16 572)

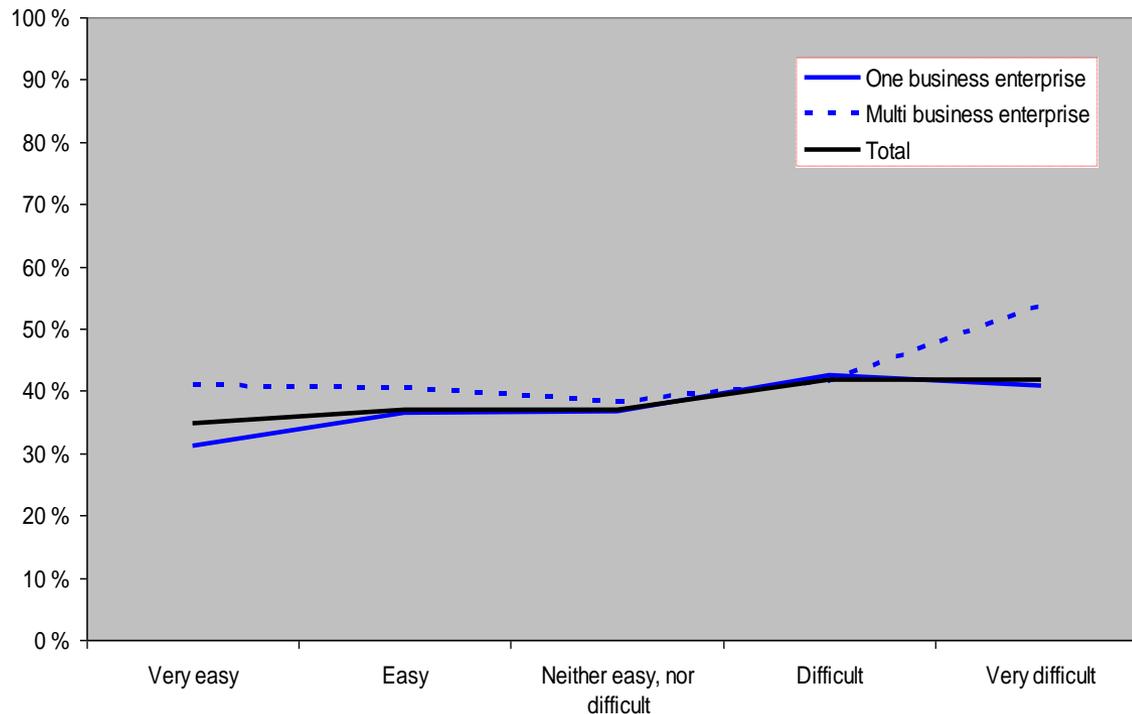
Perceived Response Burden



Sources of Response Burden



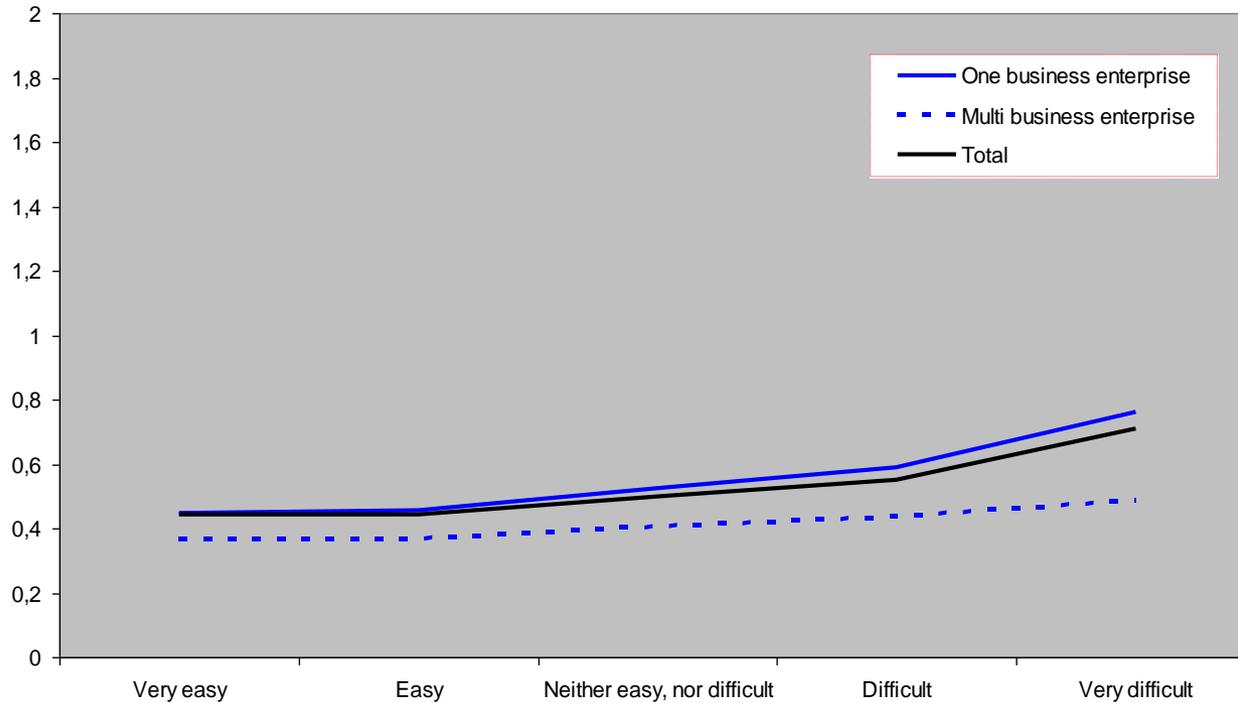
Case B (I) Proportion of questionnaires where key variable was *manually corrected* by perceived response burden



Connection, although weak



Case B (II) Average *number of reminders* by perceived response burden



Connection,
although weak



Case C. Cost deduction of production equipment statistics - Indicators

Response burden

- Time spent on reporting
- Perceived response burden
- Reporting according to deadline

Data quality

- Internal consistency
- Complete reporting
- Obvious errors



Case C

Respondents who spends much time on reporting makes more errors than those who spends little time?

Respondents who finds the reporting complex makes more errors than those who do not?

Respondents who report late makes more errors than those who report early?



Case C. Results - Data quality by completion time, perceived response burden and acceptance of deadline

| | Completion time | | Perceived response burden | | Acceptance of deadline | |
|----------------------|-----------------|-----------|---------------------------|-----------|------------------------|-------|
| | Effective | Consuming | Easy | Difficult | Before | After |
| Internal consistency | 90,91 | 93,24 | 92,19 | 93,27 | 92,80 | 93,02 |
| Complete reporting | 95,45 | 87,84 | 93,75 | 88,46 | 90,40 | 90,70 |
| Correct reporting | 86,36 | 95,95 | 92,19 | 93,27 | 92,00 | 95,35 |

→ Weak connection, in varying directions

Conclusions I

- The correlation between response burden and data quality (still) seems to point in different directions...
- Case A: No significant connection between perceived response burden and control violations
- Case B: Weak connection between perceived response burden and both manual edits and number of reminders
- Case C: Weak connection in unexpected direction between response burden and missing, consistency checks and errors



Conclusions II

Do the observed results reflect the actual situation?

High response burden:

- difficulty with understanding leads to (both high PRB and) high response error
- conscientiousness could lead to (high PRB but) low response error

Low response burden:

- sloppy attitude (or just underestimating complexity) leads to both low PRB and to high response error
- completing questionnaire without any problems causes low PRB and low response error



Conclusions III

The connection is dependent on

- * The reason for response burden
- * How we operationalize “data quality” and “response burden”
- * The connection could be hidden by background variables both affecting perceived response burden and response error

In general

- * There is no correlation between late reporting and (increased) time consumption
- * Those who report late, tend to find the reporting both easier and more complex than those who report early
- * Those who spends the least amount of time, find the reporting easy



Now what?

→ At the end of 2015 we will have an new dataset from structural statistics.

→ We will continue to advocate for a “shift” in how we understand response burden –
actual response burden = perceived response burden

→ We have established a connection, although weak and dependent on many variables – the next task will be to examine specific response burden variables in order to determine which factors are mostly influencing data quality



Thanks for your attention!



Statistisk sentralbyrå
Statistics Norway

Kari-Anne Lund and Gustav Haraldsen
Department of Data Collection and Methods, Statistics Norway