Effect of cross validation in online questionnaires on subsequent data editing - improving data quality in business surveys for National Statistics

Hanne-Pernille Stax, Peter Tibert Stoltze & Helene Feveile[1]

Statistics Denmark

Abstract

In online questionnaires we have the opportunity to give instant feedback to the respondent about conspicuous values entered. The respondent can then review and edit or confirm the conspicuous data, thus minimizing the risk of error, subsequent error-upon-error and the need for re-contact. In theory this should improve data quality and reduce burden upon the respondent at the same time. In this paper two business surveys carried out by Statistics Denmark are examined, namely Transportation of goods by lorry and Number of vacant positions. The questionnaire designs are explained, including implemented cross validations between known and new values. The different types of evoked responses to conspicuous data are demonstrated: From notification and warnings that can be ignored, to errors that must be corrected. The quality of the data is assessed by the amount of conspicuous values found in the submitted data - before and after implementation of validation in the online questionnaires. Finally, perspectives for future investigations along the same lines are drafted, including the use of para-data in the subsequent estimation process.

1. Introduction

Statistics Denmark initiated the process of converting traditional paper questionnaires for business surveys into web questionnaires for online completion in 2008. Initially the objective was to achieve faster and cheaper data collection. But the digital mode offers new possibilities for supporting the respondents and enhancing data quality by implementing responsive design and immediate micro data validation during completion of the online questionnaires.

[1] Hanne-Pernille Stax, PhD, Deputy Head of Division, Business Surveys, Statistics Denmark, hps@dst.dk
Peter Tibert Stoltze, Deputy Head of Division, Research and Methods, Statistics Denmark, psl@dst.dk
Helene Birgitte Feveile, Head of Section, Research and Methods, Statistics Denmark, hfe@dst.dk

The implementation of responsive design features in web questionnaires has largely been guided by increasing technological capability and by demand from the respondents who expect immediate response if they enter conspicuous data, so that they may review and correct or confirm before submission. Very few studies have been conducted to follow up and document the actual effect of the implementation of specific responsive design features on data quality and response burden in surveys for national statistics. In this paper we will illustrate the gradual implementation of responsive design features in questionnaires for business statistics in Statistics Denmark, and how they have affected the need for re contact and data editing of the submitted data. We will also address current challenges with regard to the need to rethink the data editing process when micro data editing is to be executed during - rather than after - data collection.

2. From "flat" digital copies to responsive web questionnaires with cross validation

In the first wave of web questionnaire design for business surveys at Statistics Denmark, web questionnaires were designed as digital copies for the existing paper questionnaires. Only minimal restrictions and responsive data checks were built into the web questionnaires, as we did not wish to burden the respondent with error messages and hard stops. Data editing was performed *after* the data collection was completed and respondents were re-contacted in order to correct or explain possible errors in the submitted data.

From user tests and questionnaire evaluations we learnt that respondents to business surveys expect dynamic validation of the entered data before they submit a web questionnaire. In response to this demand, and utilizing new technological possibilities in the second wave of web questionnaire design, part of the data editing process, which has conventionally been carried out after data collection, was moved into the web questionnaire. The questionnaires were designed with built in skipping patterns and responsive validation of entered data with regard to data type, value range, missing values etc., generating instant feedback to the respondent as data was typed into the questionnaire. Respondents were presented with assisting warnings or hard stops, and encouraged to check, correct, confirm or explain conspicuous values before submitting the questionnaire.

Moving into a third wave of web questionnaire design, still more advanced validation mechanisms may be built into the questionnaires. When possible, entered values are immediately compared to relevant other values which have been entered in the same questionnaire - or compared to other known values gathered from other sources and prefilled to the questionnaire for a specific unit.

Online cross validation of micro data is promising, but it is a delicate business. Mismatch between known and new entered values might result from error in the entered OR the known values. And cross validation with erroneous data from other sources might result in extra burden and in reduced rather that improved data quality. Thus cross validation must be implemented with caution, and follow up analysis of the effect on the quality of submitted data is crucial. Actual implemented cross validation and other responsive design features and their effect on data quality in two business surveys for national statistics is described in the following.

3. Micro data cross validation in web questionnaire for Transportation of goods by lorry

The objective of the survey is to monitor volume and variation in lorry transportation in tonne km. For each specific truck in the sample each individual trip driven in a specified reference week must be reported, entering information on length of trip and weight and type of goods. To facilitate data editing the respondent must also enter the total amount of kilometres driven by the truck in the specified reference week, and the area code of starting and end point of each individual trip.

Responsive design was not used in the first digital version of the questionnaire for the transportation survey from 2009. Cross validation of the submitted data *after* data collection indicated low data quality: Individual reported trips were not linked, empty trips seemed to be missing and the reported length of the specified trips were unreliable: In average the sum of individually specified trips equalled 2 x the reported total amount of kilometres driven in the reference week.

3.1 Redesign of web questionnaire with responsive edit checks

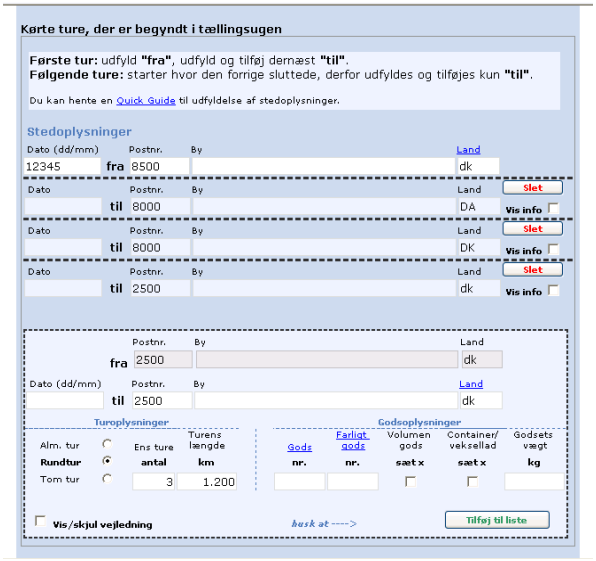
In the redesign of the web questionnaire from 2011 it was decided not to implement hard stops in order not to discourage the respondents from using the web questionnaire. Instead a number of "soft" responsive design features were implemented in order to assist and support the respondent in completing the response task as intended.

Total amount of kilometres driven during the reference week is collected as the first information in the web questionnaire – calculated from entered values from the km counter. Total length of repeated trips along the same route is dynamically calculated and length is exposed in the list of all reported trips. Lengths of reported trips are automatically summed and the sum exposed directly below the list of trips. The entered total of km driven is copied from the top of the questionnaire and exposed directly below the sum of the individual trips – for cross reference. The font colour of the sum of individual trips change from black to red, if the sum exceeds total km driven during reference week.

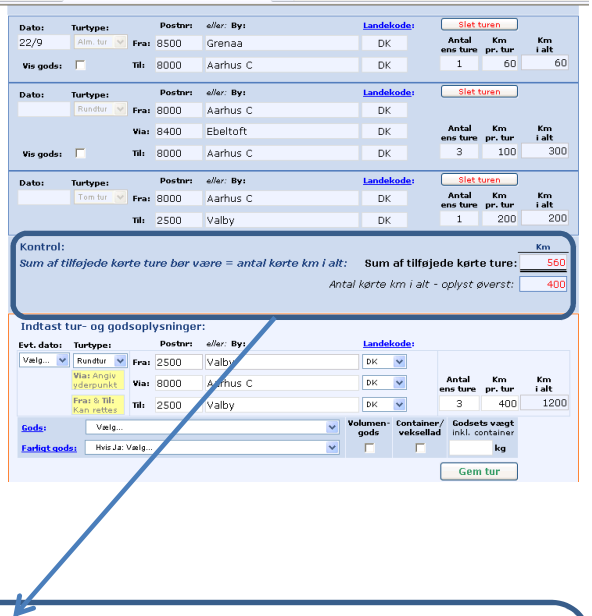
The entered end point of a reported trip (area code and city) is automatically transferred and exposed as the starting point of the following trip. The respondent is allowed to correct the prefilled starting point, e.g. if the truck has been used for private purposes, has been transported by ferry ea. A number of text fields were substituted by drop down lists offering all valid response options – and none other.

Fig. 1

2009 version of web questionnaire



2011 version of web questionnaire



Control:
Sum of entered individual trips should equal total amount of kilometers driven:

Sum of individual trips: _____
Total amount of kilometers driven – as reported above: _____

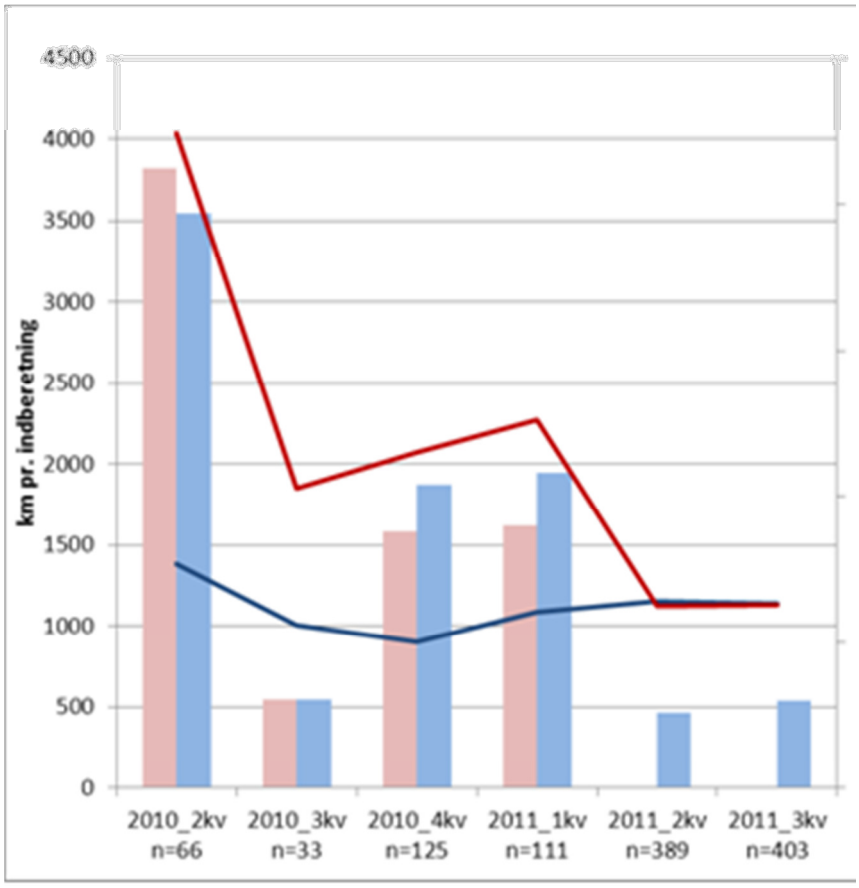
3.2 Effect of responsive design

The redesign of the web questionnaire with responsive features has resulted in a high level of linked trip, inclusion of empty trips in reported trips, a low span between reported km driven in total and sum of individual entered trips and no series break in data with regard to km driven in total.

Fig 2. Average total and average sum of trips per rapport before and after redesign I 2011-Q2

Version	Reference Quarter	Rapports Total	Rapports by web	Average km in total pr. Rapport	Average sum of trips pr. Rapport
1	2010-2	1698	70	1382	4034
1	2010-3	1707	33	1003	1848
1	2010-4	1606	135	899	2075
1	2011-1	1619	121	1089	2669
2	2011-2	1353	389	1149	1125
2	2011-3	1366	404	1141	1131

Fig. 3: Average total and average sum of trips per rapport before and after redesign I 2011-Q2



4. Responsive cross validation of micro data in web questionnaire for Vacant positions

The objective of the survey is to monitor number of job vacancies and job vacancy rates by industry, unit and size. For each work unit in the sample number of vacant positions and number of employees at a specific date in the reference month must be reported.

Responsive design features were not used in the first digital version of the questionnaire for the vacant positions survey from 2009. Data editing *after* data collection did require substantial re contact to responding units, regarding entered 0 in number of employees or regarding entered number of employees which were conspicuously high in comparison to the registered number of employees at the unit according to the business register. A conspicuously high reported number of employees indicate that the report does not cover the intended work unit but the entire legal unit.

4.1 Redesign of web questionnaire with responsive edit checks

In the redesigned questionnaire from 2011-Q3 responsive validation on number of employees was implemented to check and notify the respondent immediately, if an entered number of employees is

conspicuously high, indicating that the report is being made on the basis of a wrong unit. Number of employees according to the business register is prefilled to the web questionnaire for each specific work unit and entered number of employees is compared to the data from the business register. If the work unit has participated in the survey for a year or more, the reported – and error previously checked - number of employees 1 year back in time is also prefilled to the questionnaire and used in cross validation of the entered number of employees. The prefilled numbers of employees from the two sources may not be identical, and thus they are not displayed in the web questionnaire in order not to confuse and implement extra burden.

A warning message and request to check, correct or explain and confirm the entered number of employees is presented in the web questionnaire if a respondent enters 0 for number of employees or if entered number of employees is conspicuously high compared to the *both* prefill values for the unit. The extra comparison with the number of employees reported – and error checked - one year back in time is implemented in order not to burden respondents with error messages in the case where there might be an error in the business register value.

Fig 4.

Number of employees at work unit (control variable)
Number of vacant positions at work unit (core variable)

Hidden unit prefill: Number of employees
- from business register
- from previous survey ($y_{t\text{-}4}$)

Warning if entered value differs too much from unit prefill values (indication of wrong unit):
Number of employees at work unit "xyz" seems high. Please check, correct or explain and confirm □

4.2 Method of analysis

In the analysis of possible error in the web survey data before and after implemented responsive validation y_t is the reported number of employees at a specific date in a reference quarter t. For most units the number of employees according to the business register y_{BR} is also available. For a

subsection of the units we also have access to the variable y_{t-4}: The number of employees reported four quarters back in time. Finally, an additional indicator variable z is available, stating if a comment has been submitted by the responding unit ($z = 1$) or not ($z = 0$). After redesign a comment is requested if a conspicuous value is entered, e.g. if reported number of employees (y_t) is 0 or is conspicuously high compared to y_{BR} or y_{t-4}. The comments are used as qualitative data in the editing process and have reportedly contributed to a substantial decrease in re-contact to respondents.

A total of six error checks resembling linear edits have been made. However, not all checks are applicable to all observations. E.g. if the unit did not participate in the survey at t-4 then a comparison between y_t and y_{t-4} is not possible. Note that y_t is never missing.

Fig. 5

Error	Error check	Error flagged if	Applicable if
# 1	Number of employees = 0	$y_t = 0$	All observations
# 2	Number of employees = 0 and no explanation	$y_t = 0 \wedge z = 0$	All observations
# 4	Number of employees > 2 x business register value for unit (unit size in BR $\geq$ 50)	$y_t > 2y_{BR}$	$y_{BR} \geq 50$
# 5	Number of employees > business register value + 50 (unit size in BR < 50)	$y_t > y_{BR} + 50$	$0 < y_{BR} < 50$
# 6	Error 4 and number of employees > 2 x survey value one year back (unit size in BR $\geq$ 50)	$y_t > 2y_{BR} \wedge y_t > 2y_{t-4}$	$y_{BR} \geq 50 \wedge y_{t-4} \neq .$
# 7	Error 5 and number of employees > survey value one year back + 50 (unit size in BR < 50)	$y_t > y_{BR} + 50 \wedge y_t > y_{t-4} + 50$	$0 < y_{BR} < 50 \wedge y_{t-4} \neq .$

Error check #1: 0 employees reported is marked as a possible error. A unit must generally have at least 1 employee also counting the owner. #2 allows for zero employees if a comment is submitted.

Error checks #4 and #5 compare reported number of employees y_t with unit size according to the business register y_{BR}. For larger units ($y_{BR} > 50$) y_t must not exceed the register size by more than a factor 2.0 - which is a wide margin. For smaller units y_t must not exceed the register size by more than +50, which could be an even wider margin (eg. $y_{BR} = 20$ and $y_t = 60$ is not considered an error).

Error checks #6 and #7 build upon #4 and #5: If reported number of employees y_t would be marked as a possible error in comparison with y_{BR} but *not* in comparison with a previously reported number y_{t-4} , then it is not marked as an error. Instead it is a strong indication of an error in the business register - or a difference in perception of the business unit being inquired.

4.3 Effect of responsive design

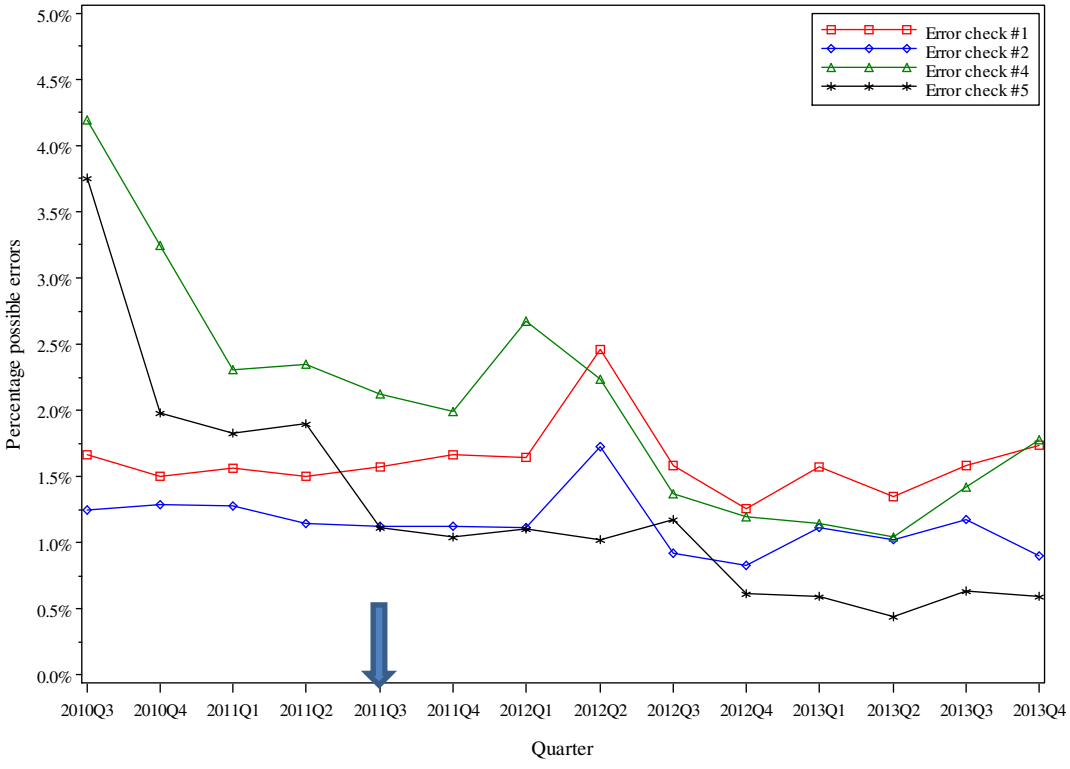
Fig. 6 and 7 shows results of error checking in web survey data during 14 quarters. Number of checks performed (n) and errors found (m) is reported for each error type. The rate of possible errors

in percent is calculated as the number of errors marked relative to the number of checks performed (*m*/*n*). E.g. in 2010-Q3 error check #1 was performed for 2163 observations and 36 observations marked as possible errors. Hence the rate of possible errors is 1.7 percent. The error checks 6, and 7 gives rise to very few possible errors being marked and are not being showed graphically in fig 7.

Fig. 6: Result of error checking

	Error check #1			Error check #2			Error check #4			Error check #5			Error check #6			Error check #7		
	N	Sum	Mean	N	Sum	Mean	N	Sum	Mean	N	Sum	Mean	N	Sum	Mean	N	Sum	Mean
Kvartal (dato)																		
2010Q3	2163	36	1.7%	2163	27	1.2%	1121	47	4.2%	1039	39	3.8%	0	.	.	0	.	.
2010Q4	3189	48	1.5%	3189	41	1.3%	1666	54	3.2%	1519	30	2.0%	0	.	.	0	.	.
2011Q1	3200	50	1.6%	3200	41	1.3%	1603	37	2.3%	1590	29	1.8%	249	1	0.4%	189	2	1.1%
2011Q2	3588	54	1.5%	3588	41	1.1%	1793	42	2.3%	1794	34	1.9%	317	2	0.6%	249	4	1.6%
2011Q3	4002	63	1.6%	4002	45	1.1%	1932	41	2.1%	2067	23	1.1%	1520	7	0.5%	1174	7	0.6%
2011Q4	4255	71	1.7%	4255	48	1.1%	1858	37	2.0%	2395	25	1.0%	1507	8	0.5%	1167	6	0.5%
2012Q1	4387	72	1.6%	4387	49	1.1%	1946	52	2.7%	2439	27	1.1%	1498	8	0.5%	1153	1	0.1%
2012Q2	4755	117	2.5%	4755	82	1.7%	2017	45	2.2%	2738	28	1.0%	1573	5	0.3%	1492	5	0.3%
2012Q3	4036	64	1.6%	4036	37	0.9%	1903	26	1.4%	2133	25	1.2%	1315	3	0.2%	1224	5	0.4%
2012Q4	3972	50	1.3%	3972	33	0.8%	1845	22	1.2%	2127	13	0.6%	1322	3	0.2%	1346	1	0.1%
2013Q1	3880	61	1.6%	3880	43	1.1%	1844	21	1.1%	2036	12	0.6%	1330	4	0.3%	1317	0	0.0%
2013Q2	4294	58	1.4%	4294	44	1.0%	2014	21	1.0%	2279	10	0.4%	1459	3	0.2%	1334	0	0.0%
2013Q3	4431	70	1.6%	4431	52	1.2%	2048	29	1.4%	2382	15	0.6%	1657	3	0.2%	1480	2	0.1%
2013Q4	4787	83	1.7%	4787	43	0.9%	2084	37	1.8%	2702	16	0.6%	1661	9	0.5%	1560	0	0.0%

Fig. 7: Rates of possible errors over time for error checks 1, 2, 4 and 5 Note that the figure displays web data only



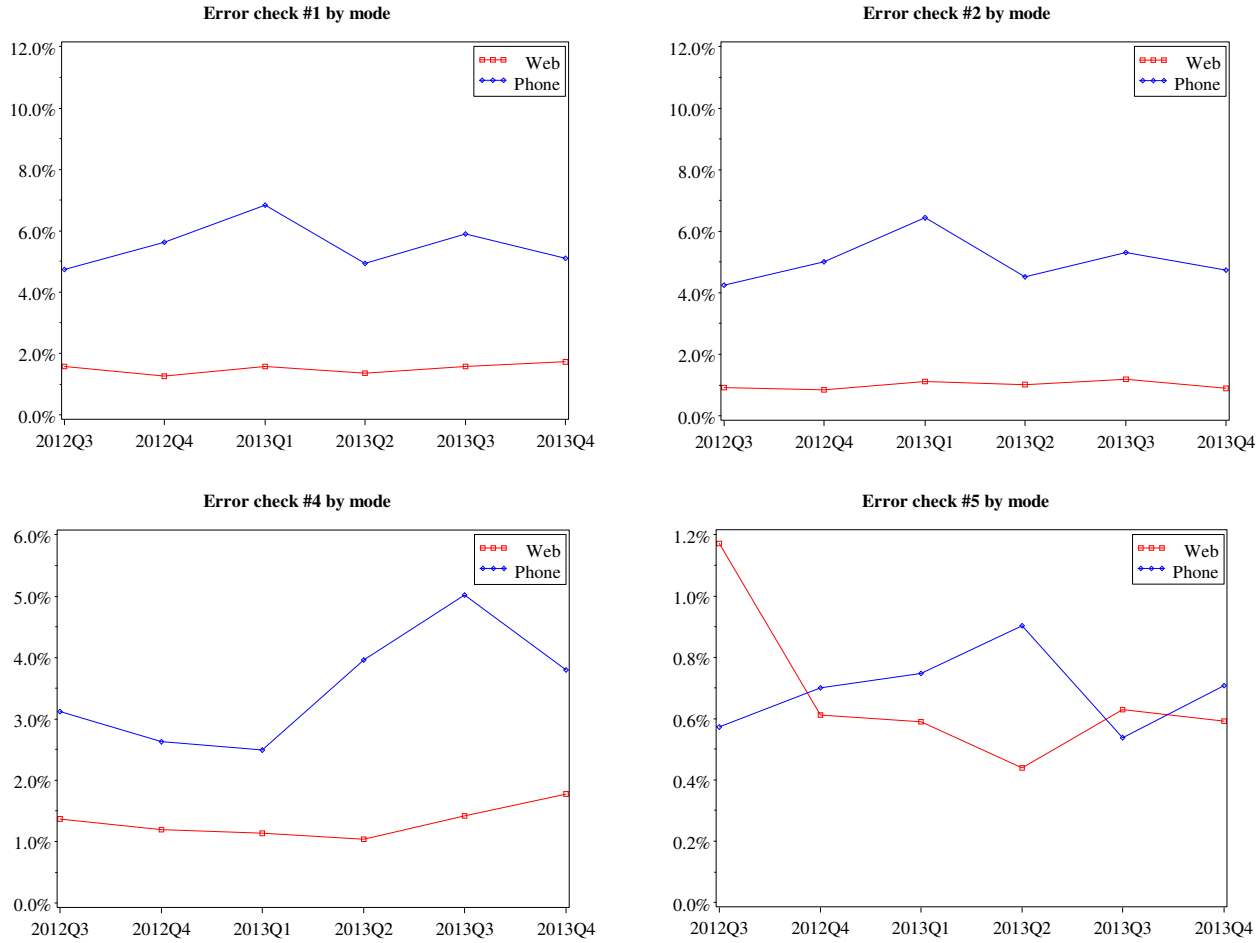
The error checks 1 and 2 (zero employees reported) shows a relatively stable level over time with the possible exception of 2012Q2, where new units were added to the sample. The offset between the two is explained by error # 2 being a subset of error # 1. No significant development in the deviation between the two is seen as a result of the redesign.

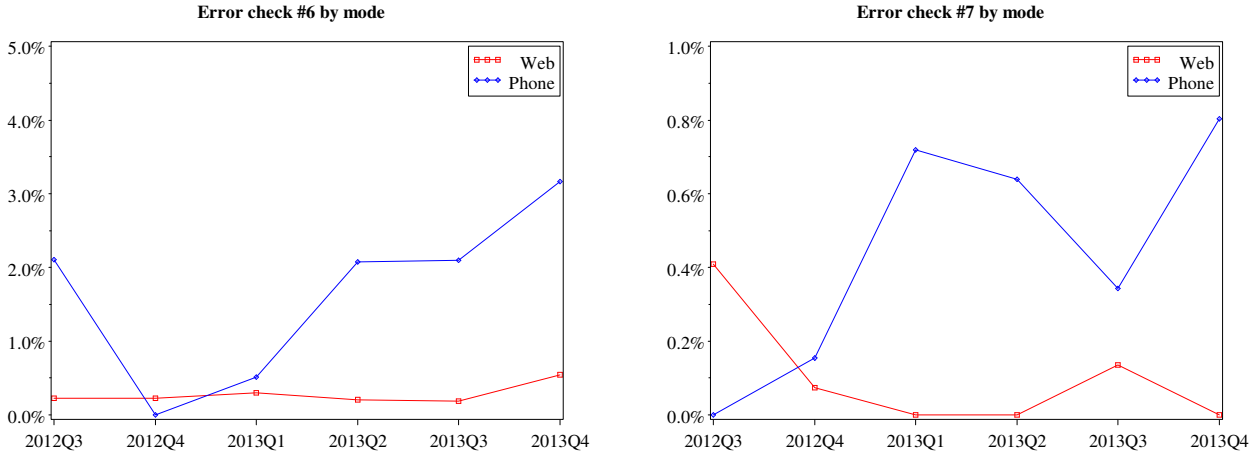
Possible errors 4 and 5 (high reported number of employees relative to unit size in business register) are seen to decrease over time. With the possible exception of error check # 5, no significant effect of the implemented validation is seen between 2011-Q2 and 2011-Q3. When the marking of a possible error is seconded by comparison with previously reported number of employees, as by error checks 6 and 7, the number of possible errors is very low (less than 10 for each quarter).

4.4. Comparison with data from key telephone reporting solution

Data collection by key telephone was introduced from 2012-Q2. Fig. 8 compares the rate of possible errors between the two collection modes, i.e. web and key telephone. Generally, the rate of possible errors is higher for data collected by telephone compared to data collected by web.

Fig. 8. The rate of possible errors over time for error checks 1, 2, 4, 5, 6 & 7 by web and key telephone.





Error checks 1 and 2 (zero employees reported) both show lower levels for data collected by web than by key telephone. Possible error 4 (high reported number of employees relative to unit size in business register - for large units) also shows lower levels for web data than for telephone data. Error check 6 (high reported number of employees relative to previously reported number - for large units) shows an unclear pattern. Error checks 5 and 7 (high reported number of employees for small units) shows low levels for both modes.

5. Conclusion and perspectives

The analysis of error level in the Transportation survey data after implemented responsive cross validation documents, that fairly simple responsive edit checks may go a long way with regard to improving data quality. Thus responsive validation may be promising, if you hit the right level. The effect of the rather more complex cross validation in the Vacant positions survey is less evident. The web questionnaire with cross validation renders better data with less possible errors than the key telephone reporting solution - and more so for large units. But there is no clear evidence of the data quality being a direct result of the implemented cross validation.

As mentioned above, the implemented responsive error checks for conspicuous unit size has rather wide margins, suitable for identification of obvious errors in conventional error checking *after* data collection. In the web questionnaire the margins might be tuned more tightly, so that lesser discrepancies between entered and known values would generate a warning and request to check and correct or explain. Optimizing the margins for responsive edit checks in web questionnaires is a delicate business, since there is no way back to “the raw data”, once an edit check has been implemented during data collection. The collection of para data on rate of activated error checks and rate of corrected values during completion may be valuable input to a more finely tuned responsive validation in web questionnaires.