Measuring data quality by the use of a routine re-interview module, some experiences from the Norwegian European Social Survey.

Kleven, Øyvin
Division for Data Collection Methods, Statistics Norway
Akersveien 26,
Oslo, Norway
E-mail: oyvin.kleven@ssb.no

Berglund, Frode
Division for Data Collection Methods, Statistics Norway
Akersveien 26,
Oslo, Norway
E-mail: frode.berglund@ssb.no

Introduction

Non-sampling errors in household surveys have received considerable attention in the past decades, as these errors clearly have become more and more critical to the accuracy of survey based statistics. As stated in the ESS Handbook for Quality Reports (p.32): A purpose of official statistics is to produce estimates of unknown values of quantifiable characteristics of a target population. Estimates are not equal to the true values because of variability (the statistics change from implementation to implementation of the statistical process due to random effects) and bias (the average of the possible values of statistics from implementation to implementation is not equal to the true value due to systematic effects; the bias of an estimator equals the difference between its expected value and the true value). It is common to separate between sampling errors and non-sampling errors. *Sampling errors*, which apply only to sample surveys; are due to the fact that only a subset of the population is selected. *Non-sampling errors*, which apply to all statistical processes is often categorised as coverage errors, non-response errors, processing errors and measurement errors.[1] Measurement errors are errors that occur during data collection and cause the recorded values of variables to be different from the true ones. Their causes are commonly categorized as (ESS Handbook for Quality Reports: 44) *Survey instrument*: the form, questionnaire or measuring device used for data collection may lead to the recording of wrong values. *Respondent*: respondents may, consciously or unconsciously, give erroneous data; *Interviewer*: interviewers may influence the answers given by respondents. Measurement errors may be

[1] Coverage errors (or frame errors) are due to divergences between the frame population and the target population. Nonresponse is the failure of a sample survey (or a census) to collect data for all data items in the survey questionnaire from all the population units designated for data collection. Nonresponse error is the difference between the statistics computed from the collected data and those that would be computed if there where no missing values. Between data collection and the beginning of statistical analysis, data must undergo processing comprising data entry, data editing, often coding and imputation. Errors introduced in these stages are called processing errors.

difficult to detect unless they lead to illogical or inconsistent responses. For many surveys, measurement errors are the most damaging source of error (Biemer & Lyberg 2003). It is well documented in the text books on survey methodology that there are many pitfalls in obtaining an accurate response from a survey question (E.g Groves et al. 2009). One approach in detecting measurement problems in surveys is to re-test some of the questions on the same sample in the same survey. In the European Social Survey (an academically driven survey who runs every second year) there is a requirement that some of the questions are asked again literary or with a slightly different answering format to the same respondents. We want to demonstrate that this approach is very useful when dealing with data quality and measurement errors.

Assessing measurement error

Measurement error can be systematic or random. Random errors are often associated with the idea of replication, i.e., if the measurement process is repeated many times from the same unit under fixed conditions the registered measurement values will vary randomly whereas the systematic error will stay constant. Response to a survey question involves a cognitive process, including comprehension of the question, retrieval of relevant information, use of that information to make required judgments and selection and reporting of an answer (Tourangeau, Rips and Rasinski (2000). Studies have shown that there are many pitfalls in obtaining an accurate response from a survey question. Groves *et al.* (2009) e.g., listed seven problems that can lead to measurement errors in a survey: (1) failure to encode the information sought, (2) misinterpretation of the questions, (3) forgetting and other memory problems, (4) flawed judgment or estimation problems, (5) problems in formatting an answer, (6) more or less deliberate misreporting, and (7) failure to follow instructions. It is easy to describe different sources in the survey process that can cause measurement errors, it is much more difficult to quantify measurement errors in surveys statistics. In the survey research literature there are some methods and techniques described that can help us understand and asses measurement errors (Table 1). The best way to deal with measurement errors is to try to prevent them. We do that by applying the best known methods where we believe measurement errors will be low, we try to eliminate measurement errors in the planning process of the survey where we do expert reviews in order to identify problems with the questionnaire based on previous research and theory on questionnaire design. Although there are many excellent textbook on how to construct questionnaires, there is no grand theory on how to construct questionnaires that eliminates measurement errors on every survey variable (e.g., Sudman and Bradburn 1982).

Table 1. Methods and techniques for assessing measurement errors in surveys

Method	Purpose	Limitations
Expert review of questionnaire	Identify problems with questionnaire layout, format, question wording, question order, and instructions	No grand theory exist who can prescribe best practice in every case
Cognitive lab methods - Behaviour coding - Cognitive interviewing	Evaluate one or more stages of the response process	Results could be biased Detecting a problem in not the same as fixing it
Debriefings - Interviewer group discussions - Respondent focus groups	Evaluate questionnaire and data collection procedures	Results could be biased
Observations - Supervision observation - Telephone monitoring - Recording	Evaluate interviewer performance Identify questionnaire problems	Costly Detecting a problem in not the same as fixing it
Split ballots experiments	Estimate/asses bias in survey estimates	Need external validation
Re-interviews (repeated interviews with the same respondent)	Estimate reliability in survey estimates	Increase costs and response burden Contexts effects, memory effects
Record check	Estimate/asses bias and/or reliability in survey estimates	Possible for only a few survey variables
Internal consistency	Estimate reliability in survey estimates	No external validation

Based on Biemer and Lyberg 2003: 261

We use cognitive lab methods like cognitive interviewing where normally 5 – 25 potential respondents are exposed to the survey instrument. By applying this method we often discover problems with single question formulation, answer categories etc. Cognitive lab methods is excellent to discover problems, because we have reason to believe that when a small sample of the respondents have problems with a question also many other respondents have similar problems. However, these methods do not give us necessary any information about the prevalence of the problem, and pointing at a problem is not the same as to fixing the problem. We are often faced with the situation that it is almost impossible to know beforehand how to formulate a question with low measurement error.

Measurement theory in psychology, called psychometrics, splits measurement errors in terms of validity and reliability. The validity of a measurement refers to the extent to which the measurement accomplishes the purpose for which it is intended (Alwin, 2007:22). Validity (construct validity) is almost impossible to measures directly. Reliability is about the consistency of the measurement at hand. Do respondents deliver consistent responses when the same question is repeated within a short time? If a question is reliable, it should produce the same response from the same respondent as long the time lag is so short that one can not suspect that a real change not has taken place. One way to evaluate the reliability of a question is to investigate the consistency

of the responses by repeated measures. That is what is done in *classical true-score theory* (CTST) (Alwin, 2007:35). The true-score is unknown – and may not exist in "reality," only in the model. However, if the repeated measures are reliable, one can assume that these measures reflect true-score. Consequently, if a measure is reliable, repeated measures of it will have a high level of correspondence. Reliability represents a crucial aspect of the quality of the data, and is essential in judging the *validity* of a (quantitative) measure. The measurement itself is an indicator of a theoretical concept, and measurement validity expresses a concern with the linkage between concept and the indicators. If that linkage is good, it is necessary that the reliability has to be high. In order to be judged as a valid measure, responses on the same question can not be given at random. There has to be a certain level of correspondence between the two responses from the same respondent. So, reliability is indeed a necessary condition for the validity of a measure, which makes it an important topic to investigate. On the other hand, reliability is not sufficient to conclude that a measure is valid. A wrongly adjusted measure will repeatedly produce reliable data, even if the measure is not right.

Measuring reliability in reinterviews by raw agreement rates

The European Social Survey is an academically driven cross-national survey that has been conducted every two years across Europe since 2001 (http://www.europeansocialsurvey.org/). The survey measures the attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations. Any measurement will contain errors and these problems can be even larger in comparative research because the differences in measurement errors can cause differences across countries which have nothing to do with substantial differences. In the survey it is a standard procedure in every round that a supplementary questionnaire is used. The net sample from the survey is randomly assigned to three different groups who are asked some of the questions again. At least one or two groups are given the same question in a way that makes it possible to estimate the reliability of the question.[2] In this presentation we limit our analysis to two single questions from the 2008 edition in Norway: '*self-placement along the left-right scale*' and '*satisfaction with national government'*. Data from the main questionnaire was administered by face-to-face interview, and data from the supplementary questionnaire was administered as a self-completion

[2] The supplementary questionnaire is a separate questionnaire that makes up part of the core module. It is administered after the socio demographic questions and the rotating modules. It is designed to evaluate the reliability and validity of items in the main questionnaire using the Multi-Trait Multi-Method (MTMM) approach. The data from all the different rounds and from the different countries is analysed at the Research and Expertise Centre for Survey Methodology at the University Pompeu Fabra in Barcelona, with the aid of structural equation models (LISREL) (see Saris & Gallhofer 2007, Saris et al. 2010, http://sqp.upf.edu/)..

questionnaire. Reinterviews can be used to asses' reliability with the following assumptions: (1) There are no changes in the underlying construct between the two interviews. (2) All the important aspects of the measurement protocol remain the same ("the essential survey conditions" remains the same). (3) There is no impact of the first measurement on the second responses (there are no memory effects; the second measurement is independent of the first) (Groves et al 2009:282). In reality these assumptions are hard live up to, but we will argue that in this case it is not likely that the difference in administration mode will have a huge impact on the responses. Context effects and memory effects can of course be present. In order to minimize the memory effect it is necessary to maximize the time period between the two questions. In our case it is at least on hour between the two measures. Van Meaurs and Saris (1990) suggest that at least 20 minutes are required as time period between two measures. Saris et al (2010:65) cites a number of laboratory studies, and claiming that people cannot remember attitude reports they made one hour previous.

Table 2. Re-interviewing design for left and right self placement

	Question	**Answer categories**
Main Questionnaire (f2)	In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right?	Horizontal 11 point scale only labelled at the end points (Left=0, Right=10)
Version 1 (Self completion)	In politics people sometimes talk of "left" and "right". Where would you place yourself on this scale? **Please tick one box.**	Horizontal 11 point scale only labelled at the end points (Left=0, Right=10)
Version 2 (Self completion)	In politics people sometimes talk of "left" and "right". Where would you place yourself on this scale? **Please tick one box.**	Horizontal 11 point scale only labelled at the end points (Extreme Left=0, Extreme Right=10)

In this study we use a very simple method to assess the reliability between two measures of the same item (question). We simple cross tabulate the two measures, and count the number of respondents who give the same answer, agreement rate. We also look at the structure of distribution in the table. Before we analyse the data for each two measures we decide on what we call "acceptable agreement rate". This has to be decided based on knowledge based on the subject studied, not just on a technical method. Some movements in the distribution are clearly more damaging to the reliability than others. For instance if there is a middle point in an answer format to a question it could be argued that it is more damaging to the reliability if many respondents crosses that middle point. We then display this by marking in red what we consider an unacceptable shift between the two measures. Table 3 enables us to calculate the test-retest

reliability of the left right scale. The same question; exactly same wording, has been asked to the same respondents two times. Correlation coefficients, or *r* values are often recommended to compare the two sets of responses, and in general *r* values are considered good if they equal or exceed 0,70 (Litwin 1995:8). A regular Pearson correlation coefficient (*r* value) can miss the structure of the distribution. In our experience it is often also more intuitive to survey practitioners (employs in the statistical production) to just present the raw data and display the reliability in percentages. If agreement between the two measures displayed in Table was perfect (every respondent gave the same answer two times) all responses would be along the diagonal, then the raw agreement rate would be 100 per cent. In surveys we know that this is very seldom the case due to different systematic or random measurement errors. When surveying reel people about objects in the social world test-retest reliability is almost newer perfect. We as statisticians need to decide what the acceptable reliability should be. In theory the agreement rate can be zero, if no respondent give the same answer two times, this is also very unlikely. In our first example (table 3) the raw agreement rate is 64 per cent, which is probably not very impressive. In contrast the *r* value of table 3 is 0,89, which to some could indicate a very high reliability (close to perfect). But if look at the raw data table (Table 3) we see that very few respondents shifts dramatically (e.g go from hard left to hard right) between the two measures.

Table 3. t1 In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right? / t2 In politics people sometimes talk of "left" and "right". Where would you place <u>yourself</u> on this scale? **Please tick one box.** Absolute numbers

t_1 \ t_2	0 Left	1	2	3	4	5	6	7	8	9	10 Right	n
0 Left	6	1	0	0	0	1	0	0	0	0	0	8
1	1	4	0	1	0	0	0	0	0	0	0	6
2	0	3	9	5	0	0	0	0	0	1	0	18
3	1	0	10	50	6	2	2	0	0	0	0	71
4	0	1	3	9	33	2	0	0	0	1	0	49
5	0	0	1	0	10	82	13	3	1	0	0	110
6	0	0	0	0	2	12	21	8	2	0	0	45
7	0	1	0	1	0	1	8	42	20	1	1	75
8	0	0	0	0	0	1	1	9	23	4	2	40
9	0	0	0	0	0	0	0	3	4	9	0	16
10 Right	0	0	0	0	0	1	0	0	2	1	5	9
n	8	10	23	66	51	103	46	65	52	17	8	449

This is probably because there exist no accurate "true value" on the selfplacement on the left right scale. If we allow respondents to move two places on the left-right scale in each direction, marked

white in the table, the "substantial agreement rate" would be 97 per cent. Overall the test-retest on this question shows a high reliability.

Often we are also interested in testing a slightly different version of the question, maybe with a different answering format. This is often referred to as alternative-form reliability (Litwin. 1995:13). In table 4 we have introduced the word extreme on both ends of the poles. Has this any consequences for reliability?

Table 4. t1 In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right? / t2 In politics people sometimes talk of "left" and "right". Where would you place <u>yourself</u> on this scale? **Please tick one box..** Absolute numbers

t_1	t_2	0 Extreme left	1	2	2	3	4	5	6	7	8	10 Extreme right	n
	0 Left	3	1	2	1	0	0	0	0	0	0	0	7
	1	1	6	1	0	1	1	0	0	0	0	0	10
	2	0	0	10	4	1	2	0	0	0	0	0	17
	3	0	0	4	40	12	3	1	1	0	0	0	61
	4	0	1	0	12	25	12	2	0	0	0	0	52
	5	1	0	2	7	11	70	7	2	1	0	0	101
	6	0	0	0	4	1	7	28	9	1	0	0	50
	7	0	0	0	1	0	5	18	37	12	0	1	74
	8	0	0	0	0	0	3	3	16	30	0	0	52
	9	0	0	0	0	0	0	1	0	6	5	0	12
	10 Right	0	0	0	0	0	0	0	0	5	3	3	11
	N	5	8	19	69	52	106	61	65	56	8	4	453

Compared with table, the overall picture in table 3 is essentially the same. However, the overall measurements are little lower – *r* is 0.87 while the raw agreement is 57 per cent. The "acceptable agreement rate" is now 95 per cent. This indicates that adding "extreme" to the poles, has had a slight impact on the responses. Interestingly enough it is not any substantial changes in the poles. The raw data indicates that the reliability is very high when asking respondents to place themselves on the left-right scale. This is the case even when the measure is adjusted by stating that the poles are "extreme" positions. We suggest that when trying out a new version of an existing question (single item) in a survey, it is recommended to measure both test-retest reliability and alternative form reliability. We need the test-retest reliability to benchmark the alternate-form reliability.

In our next example we show two different versions of alternate-form reliability for satisfaction with the government (table 5).

Table 5: Re-interviewing design for satisfaction with government

	Question	**Answer categories**
Main Questionnaire (f2)	Please answer using this card, where 0 means extremely dissatisfied and 10 means extremely satisfied. Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job?	Horizontal 11 point scale only labelled at the end points (extremely dissatisfied=0, 10= extremely satisfied) t=10)
Version 1 (Self completion)	Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job? Please tick one box.	Horizontal 11 point scale only labelled at the end points (dissatisfied =0, 10=satisfied)
Version 2 (Self completion)	Please indicate to what extent you agree or disagree with the statements below. 'I am satisfied with the way the government is doing its job.' Please tick one box.	Agree strongly, Agree, Neither disagree nor agree, Disagree, Disagree strongly

In the left-right example, there was a minor change by adding the term "extreme" to the poles to one of the groups. In table 6 the raw agreement rate is 43 percent (*r* value is 0,75), while the "acceptable agreement rate" is 92 percent, a relatively high reliability. Another way of thinking about reliability is how robust the question is for changes. So what happens if we change the item to an Agree/Disagree question (Likert scale), and at the same time change the order of the response set. If the question is robust, we should also expect that changing direction and apply a different scale will produce responses that to a certain extent will have the same substantial meaning (table 7). The "raw agreement" rate in table 7 between the 11-point scale and the 5-point scale is 62 per cent (0,1=5; 2,3=4; 4,5,6=3; 7,8=2; 9,10=1), and r = -0,63. The, the *r* value is lower than for 11-point scale, while the raw agreement rate is higher. Only 4 per cent of the observation is the red zone, meaning that the "acceptable agreement rate" is 96 per cent. This indicates that the measure for satisfaction with national government is reliable even when the directions are changed and the format of the question is changed.

Table 6: t1 Please answer using this card, where 0 means extremely dissatisfied and 10 means extremely satisfied. Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job? / t2 Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job? Please tick one box.

t_1 \ t_2	0 Dissatisfied	1	2	3	4	5	6	7	8	9	10 Satisfied	n
0 Extremely Dissatisfied	5	2	1	0	0	0	0	1	0	0	0	9
1	3	0	6	0	0	0	0	0	1	0	0	10
2	3	2	13	2	0	0	1	0	0	0	0	21
3	1	0	9	29	5	9	1	0	0	1	0	55
4	0	1	2	12	25	11	5	1	0	1	0	58
5	0	1	3	8	18	38	15	13	4	0	0	100
6	0	0	2	4	8	17	25	9	7	0	0	72
7	1	0	0	0	6	7	17	30	14	2	0	77
8	0	0	0	0	0	4	2	7	24	4	2	43
9	0	0	0	1	0	0	0	0	2	3	0	6
10 Extremely Satisfied	0	0	0	0	0	0	0	0	0	0	1	1
n	13	6	36	56	62	86	66	61	52	11	3	452

Table 7: t1 Please answer using this card, where 0 means extremely dissatisfied and 10 means extremely satisfied. Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job? / t2 Please indicate to what extent you agree or disagree with the statements below. 'I am satisfied with the way the government is doing its job.' Please tick one box.

	1 Agree strongly	2 Agree	3 Neither agree nor disagree	4 Disagree	5 Disagree strongly	
0 Extremely Dissatisfied	1	0	1	3	4	9
1	0	1	0	4	4	9
2	0	1	2	8	5	16
3	0	0	18	33	2	53
4	1	2	27	20	0	50
5	0	7	67	16	2	92
6	0	19	53	9	2	83
7	0	48	31	4	0	83
8	2	35	7	1	0	45
9	2	4	1	1	0	8
10 Extremely Satisfied	0	2	0	1	0	3
	6	119	207	100	19	451

Discussion

Our example in this paper is from an academically driven survey, but we want to stress that this approach will be very useful for also official statistics. Although it is well recognised in the survey research litterateur and in ESS handbooks on quality that re-interviewing is useful for

understanding and assessing measurement errors, there are to our knowledge few examples of ongoing surveys in official statistics that uses re-interviewing as part of the quality profiling of the statistics. There are of course plenty examples from pilots and pretest before we do the actual survey in official statistics, but we suggest that this could be included in the survey design on key variables. Obstacles and constraints to this approach is of course several. It will rise the cost of the survey, it will increase the responseburden, it is not possible to be 100 per cent sure that we will capture "the true" reliability, validity and bias with only two measures, there will always be some factors that we don't control (context effects, interviewer effects, memory effects etc). Many practitioners within he NSIs seems to believe that the approach need complicated and sophisticated statistical analysis. We believe that sophisticated statistical analysis can be very useful when assessing measurement errors, but often a simpler statistical method is preferable to a more complicated one. In the words of John Uebersax: *All other things being equal a simpler statistical method is preferable to a more complicated one. Very basic methods can reveal far more about agreement data than is commonly realized. For the most part, advanced methods are complements to, not substitutes for simpler methods* (http://www.john-uebersax.com/stat/raw.htm). Questionnaires tested and/or development in cognitive labs should be tested in re-interviews. Re-interviews can reveal problems in single questions, but it can also reveal that some "problems" detected in the cognitive labs are not "real" problems in survey statistics. Key variables in official statistics can be tested more extensively by re-interviewing.

REFERENCES

Alwin, D. (2007), *Margins of Error,* New York: Wiley

Biemer, P. and Lyberg, l. (2003), *Introduction to Survey Quality,* New York: Wiley

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E.,&Tourrangeau, R. (2009). *Survey Methodology*, 2nd ed. New York: Wiley.

Eurostat (2009) *ESS Handbook for Quality Reports.* Luxembourg: Office for Official Publications of the European Communities.

Litwin, Mark S (1995) *How to measure survey reliability and validity.* Sage Publikations

Saris, W.E & N. Gallhofer (2007) *Design, evaluation and analysis of questionnaires for survey research.* Hoboken: Wiley

Saris, Willem E, Jon A. Krosnick, Melanie Revilla and Eric M. Shaeffer (2010) "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Oprions" in *Survey Research Methods vol 4. No1 61-79*

Sudman, S., & Bradburn, N.M (1982). *Asking questions.* San Francisko: Jossey-Bass.

Tourangeau, R., Rips, L.& Rasinski, K. (2000). *The Psychology of Survey Response*.Cambridge:Cambridge University Press.

Van Meurs, A., & Saris, W.E (1990) "Memory effects in MTMM studies" in Saris & van Meurs *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* Amsterdam: North-Holland

Zhang, Li-Chun, Ib Thomsen and Øyvin Kleven (2013) "On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys" in *International Statistical Review* (2013)