10.1 Introduction

This chapter deals with two broad issues that are crucial in survey research: non-response and measurement error. Both rely on different research traditions and developed largely separately, although both have a prominent place in famous handbooks or conference proceedings on survey methodology (Rossi et al. 1983; Groves 1989; Biemer et al. 1991), and in the programs of international conferences such as AAPOR, ESRA, and WAPOR. Moreover, these two aspects of *Total Survey Error* (Groves 1989, pp. 8–18; Groves and Lyberg 2010) are related. Non-response and measurement error have an effect on the response distributions, and sometimes on correlations, but higher response rates do not automatically lead to less measurement error (Groves 1989, p. 218). The opposite is possible when higher response rates are obtained by including less motivated respondents in the obtained samples (Kaminska et al. 2011; Stoop et al. 2010, pp. 188–203).

Central to the Groves conception of error is the distinction between variable error (variance) and bias. The total error of a survey is the mean

square error (MSE), i.e., the sum of all variable errors (variance) and all bias. Variable error is associated with “precision”, whereas, bias relates to “accuracy”. Errors are specific to a single statistic (in a single survey) and can thus vary over the implementations of a survey design, while bias is a kind of constant error that affects a statistic in all implementations of a survey design (Groves 1989, pp. 8–14). Groves distinguishes between a survey statistics viewpoint, where (in principle) an observable true value is assumed, from the psychometric perspective wherein attributes are measured that cannot be observed by anyone (unless the respondents themselves). One can even discuss whether the attribute really exists.¹

Bias has a clear meaning in situations where an observable true value of a statistic in the population exists and an observable true value for the respondents in the sample is assumed. Bias is then the difference between these two. In this conception there is no bias in the measurement of un-observable attributes as opinions, attitudes, preferences, and values, but only variable error over repeated measurements (variance) (Groves 1989, p. 18). If one wants to measure subjective attributes one needs

J. Billiet (✉) · H. Matsuo
Centrum Voor Sociologisch Onderzoek (CeSO),
K.U. Leuven, Parkstraat 45 Bus 3601,
3000 Leuven, Belgium
e-mail: jaak.billiet@soc.kuleuven.be

H. Matsuo
e-mail: hideko.matsuo@soc.kuleuven.be

¹ See for example the discussion on so called ‘non-attitudes’ in Saris and Sniderman (2004).

observed² indicators or sets of indicators to measure the underlying unobservable characteristic (constructs or latent variables) which can only be approximated with any applied measurement. Some refer to the conception of a "true value" in this situation as a "Platonic true score" (Bomstedt 1983). In the approaches to the study of error that will be discussed further, models are proposed in which it is possible to estimate unreliability (random error) and invalidity (systematic error).

Groves distinguishes in the TSE concept two main types of errors: errors of non-observation and observational error. Errors of non-observation are coverage, non-response, and sampling. The sources of observational error are the interviewer, respondent, the research instrument, and the mode (Groves 1989, pp. 14–24). The first part of this chapter is devoted to only one aspect of non-observation error, non-response, and the second part exclusively deals with measurement error related to the research instrument (the questionnaire). This means that this chapter covers a limited part of the quality of a survey. This becomes clear when we consider the recent developed "Total Quality Management" (TQM) framework (Loosveldt et al. 2004).

TQM is developed in the context of large interview surveys in which the management of a survey by a central coordination team plays a crucial role. This is for example the case in large-scale cross-national surveys as the *European Social Survey*. This framework is based on the distinction of both a process evaluation and an output evaluation of both the obtained sample and the registered responses (measures) to questions at the other. It integrates several existing approaches as 'Total Survey Error' (Groves 1989; Groves et al. 2004), 'Continuous Quality Improvement' (CQI) (Biemer and Caspar 1994), and the 'Total Design Method' (TDM) (Dillman 1978, 2000). TQM builds on the strengths of each of these approaches to data quality assessment. Like TSE, it pays attention to the statistical error components of sample

survey statistics, but places it in a much broader perspective of data quality in accordance with the concept of 'responsive design' (Groves and Heeringa 2006) which takes costs and time into account (Biemer 2010). TQM considers, like CQI, the process of survey production, and decomposes the survey into three phases: before, during and after the fieldwork (e.g., training and fieldwork preparation, contact procedure, follow-up, and feedback). Finally, common with TDM, the TQM framework considers factors that can increase response rates by sending out cover letters, paying attention to questionnaire design, and follow-up of non-respondents through response enhancement procedures.

Data quality thus covers all stages in the research process (Lyberg et al. 1997): the design and implementation of random samples; questionnaire construction; the translation process (and its documentation) in case of multi language samples; response enhancement during the fieldwork stage; the study of measurement error of the obtained measures; the collection of events and contextual information; transparency, documentation of the data files. In order to obtain equivalent samples and responses in cross-national surveys, the whole enterprise, which is performed by several task forces and research teams, can only succeed thanks to a firm coordination of the whole. The distinction between process evaluation and output evaluation of both the obtained sample and the registered responses is certainly useful in all stages of the process of a survey production/implementation but we evidently cannot deal with all these aspects within the context of this chapter.

In the first part of this chapter on error of non-observation, some elements of process evaluation are briefly discussed. The focus is further, however, mainly on output evaluation: the assessment of the obtained sample and the study of non-response bias. For illustration purposes examples of the *European Social Survey* (ESS) are used because the quality of this cross-national survey is widely recognized³ (Jowell et al. 2007).

³ The ESS Central Coordination Team received the 2005 *European Descartes Prize* for excellence in collaborative scientific research.

² "Observed" in the meaning of "measured" in a broad sense by recording a response or by observing an event.

In the second part on measurement error, a state of the art is offered. It covers question wording effects that emerged from the early split ballot (field) experiments and the contribution of cognitive psychology to the understanding of survey response. After a short critical review of the early split ballot tradition, and some recent developments, attention is paid to the modeling of measurement error and response styles. The relation between both parts, non-response and measurement error, is briefly discussed in the conclusions.

10.2 Evaluation of Obtained Response

The first part of this chapter deals with the process and the output evaluation of the obtained sample, or response quality. It is largely based on our experience with ESS as it is reported in Stoop et al. (2010). The process evaluation covers all the planned steps of a survey production/implementation in order to optimize the response rate and to minimize non-response bias. Measures taken in developing and implementing detailed contact forms and call records which enable detailed analyses and documentation of the response process in each country; providing feedback to NCs based on the analysis of call record data in order to improve response rates in following rounds (Billiet et al. 2007a, b). The concrete implementation of these measures in ESS is documented in Stoop et al. (2010, pp. 62–63).

Not all measures are equally important in view of response enhancement. The most effective measures in ESS⁴ concern the number and timing of contact attempts and the refusal conversion strategy. This is based on the analyses of the call record data of previous ESS rounds. Evaluation measures can therefore focus on these aspects. The first objective of the contact procedure is to establish direct contact with the sample unit. It is therefore important to analyze the number of contact attempts required to establish such direct contact. Computing (ratio) scores between mean number of contact

10.2.1 Process Evaluation of Obtained Sample

With respect to the obtained sample, the process evaluation focuses on the fieldwork training and preparation, the contact procedure, and follow-up and feedback. This evaluation focuses specifically on aspects of interviewer training that deal with the planning of the visits, contact procedures, respondent persuasion, refusal conversion techniques, and the implementation of follow-up visits or quality control back-checks (Billiet et al. 2007a, b; Stoop et al. 2010, pp. 115–159).

Response enhancement measures usually target on measures occurring during the

preparation stage and the data collection stage of a survey. However, in case of successive survey, as is the case in the biennial ESS, measures taken after the fieldwork may play a role in the subsequent surveys, for example, feedback to interviewers and information about response rates according to the time of the contact attempts (Stoop et al. 2010, pp. 121–142). Response enhancement measures in the preparation stage and during the data collection stage include properly selecting and training interviewers (in doorstep techniques); sending advance letters whenever possible; extending the fieldwork period to at least 30 days in order to undertake at least four contact attempts; the timing of the contact attempts (over the weekday and time of the day); contact mode of the first contact attempt (e.g. personal visit in ESS); at least one personal visit in case of a (first) refusal by telephone; reapproaching the respondents who refuse and "conversion" attempts instead of substitution of original sampling units; using incentives when response rates remain too low; developing and implementing detailed contact forms and call records which enable detailed analyses and documentation of the response process in each country; providing feedback to NCs based on the analysis of call record data in order to improve response rates in following rounds (Billiet et al. 2007a, b). The concrete implementation of these measures in ESS is documented in Stoop et al. (2010, pp. 62–63).

Not all measures are equally important in view of response enhancement. The most effective measures in ESS⁴ concern the number and timing of contact attempts and the refusal conversion strategy. This is based on the analyses of the call record data of previous ESS rounds. Evaluation measures can therefore focus on these aspects. The first objective of the contact procedure is to establish direct contact with the sample unit. It is therefore important to analyze the number of contact attempts required to establish such direct contact. Computing (ratio) scores between mean number of contact

⁴ Discussion of this can be found in the document of ESS Round 4 specification for participating countries.

attempts and non-contact rates, and analyzing the impact of the number of contact attempts on the final response rate, is an equally important tool in view of the assessment of the effect of the response enhancement measures. Usually, response is only possible after a prior contact. Detailed documentation of the refusal conversion process in terms as initial refusal rates, proportions of reapprached refusals, and success rates of refusal conversion attempts is crucial in the context of assessment of the cost and benefit balance (Stoop et al. 2010, pp. 171–176). The increase in final response rates can be substantial depending on the combined efforts in refusal conversion activities.

10.2.2 Output Evaluation of the Obtained Sample

It is useful to take a close look at response and non-response rates since high non-response rates may increase the *likelihood* of non-response bias, although this is not unconditionally the case (Groves 2006). Correct documentation of the response, and at least equally important, the non-response, is vital for assessment of the quality of a survey. It is also indispensable for the study of non-response bias.

10.2.2.1 Documentation of the (Non-)Response Outcomes

Anybody who wants to compare response rates should understand primarily how these rates are calculated since response rates are differently computed across surveys. A standard way of computing response rates in high quality surveys these days uses the AAPOR (2006) definitions. The starting point is the basic distinction at the sample unit level between 'completed interview' (I); 'partial interview' (P); 'refusal' (R); 'non-contact' (NC); 'other' (O); and 'ineligible' (IE). The ineligible cases are excluded from the denominator of the response rate formula. In ESS, which is a cross-country survey, the ineligible cases are sampled units that are passed off or that have moved outside the country. The response rate is calculated by dividing the number of achieved valid and complete

interviews (I) by the number of eligible sample units (Stoop et al. 2010, pp. 61–67):

$$1 / (1 + P + R + NC + O)$$

Response and non-response rates can be assessed based on benchmarks, but these are mostly not the same for all single-country surveys. Effects of nonresponse in cross-nation research is even more challenging because non-response is affected by country-specific issues like, culture and past and current survey traditions (see Stoop and Harrison Chaps. 9, 15). Comparability of cross-nation data can be only achieved when data quality affected by non-response is assessed and improved (Couper and de Leeuw 2003, p. 157). Take for example that the target response was set to 70% (and a non-contact rate of 3%) in ESS. Call record data are the primary sources for the study of response rates and non-response. The files with call record data document the outcome of each contact attempt for every sample case (respondents as well as non-respondents) (Lynn et al. 2001; Blom et al. 2010). The contact forms (CF) are short and compact interviewer questionnaires applied to each sample (selected) person, and record information about the date, day, time, mode, and outcome of each contact attempt. Contact forms also contain information about the non-respondents (reason for refusal, estimation of future cooperation) and ineligible cases. Apart from the outcomes of each contact attempt, the CFs also contain observable data in the form of housing and neighborhood characteristics. In the case of address-based and household-based sample frames, the household and respondent selection procedure is noted in the contact form. All of these features in the contact form are useful for the analysis of non-response bias.

Contact record data also provide an extra opportunity for data quality checks. These data can be cross-checked against information from other sources including reports from fieldwork managers and the main file, which is the registered response file. Such cross-checks can focus on a number of indicators including the total number of sample units (response units, non-response units, and ineligible units), the total number of interviewed units by checking the consistency of information (e.g. occurrence and

time) across different data sets. In most cases, when the response rate is calculated on the basis of the contact files, the outcome of the last contact attempt is taken as the final response code. Refusals constitute an exception: even if the last contact attempt results in non-contact or another outcome, they are listed as refusals (Billiet et al. 2007a, b; Lynn et al. 2001).

10.2.2.2 The Assessment of Non-Response Bias

The problem of non-response bias has to be distinguished from the response enhancement question as such since there is no perfect relationship between the degree of non-response and the degree of bias (Groves 2006). Some scholars argue that better than aiming for high response rates one should try to minimize non-response bias by using a responsive design (Groves and Herring 2006). However, minimizing bias during fieldwork is even more difficult to implement in practical fieldwork protocols than enhancing response rates. One should realize that non-response bias varies substantially across variables (Groves 2006; Vehovar 2007). Assessing non-response bias requires either population information with respect to the core variables of a survey, or similar information on non-respondents (Stoop et al. 2010, pp. 211–277). Both are rarely available, at least in surveys on opinions, attitudes, and values. There are several operational definitions of bias (Groves 2006) that are applied in ESS.

Bias as deviation between sample and population distributions. In this definition, a comparison is made between distributions of some variables in the sample and distributions of these variables in the population, as far as these are available. This method of bias estimation has been used in the ESS (Vehovar 2007). Bias is not only estimated for documented demographic variables in a population but also, under certain assumptions, for variables that are not documented in the sample. This is done by comparing the weighted sample (according to the known variables) with the unweighted one for a large number of variables. It is clear that the success of this way of evaluation of the sample depends strongly on the correlation between the

documented variables (the so-called post-stratification variables) with the (often attitudinal) target variables in a survey. The correlations are mostly low (see Stoop et al. 2010, pp. 219–223).

Bias as difference between cooperative and reluctant respondent. This approach to study non-response bias is to compare converted refusals with cooperative respondents. Here, the underlying assumption is that with less field efforts these would have been final refusals, and with even more field efforts additional refusals could have been converted. This underlying assumption cannot always be tested and the support for it is ambiguous (Stoop et al. 2010, pp. 214–228).

Bias as difference between respondents and non-respondents. This way of studying bias more directly focuses on the characteristics of the non-respondents and on the differences between respondents and non-respondents in the gross samples. The call record data in the ESS data sets contain several possibilities. In principle, for all selected sampling units observational data has been observed about gender, age, and one can ask the interviewer to record some characteristics of the housing and neighborhood of the respondents (Stoop et al. 2010, pp. 229–241). Surveys among non-respondents some weeks after a survey is finished using a small (PEDAKS) questionnaire (Lynn 2003), as is done after ESS round 3, is a rather expensive way of obtaining additional information about bias, but is another way to study bias (Stoop et al. 2010, pp. 252–276; Matsuo et al. 2010b). Equally, using contact sequences through optimal matching as a tool for non-response adjustment purposes is another kind of approach to study this (Kreuter and Kohler 2009).

Adjustment for bias in model-based approaches. These approaches are considered: model-based estimates of response probabilities; prediction and response propensity modeling; calibration estimation; the extended use of auxiliary data at the individual level and aggregate data at (small) context level; the introducing of instrumental variables and jack-knife methods in variance estimation with replicate weights computed in one step even though the calibrations weights themselves may be determined iteratively. The information needed for applying a model-based

approach can be obtained from the sample itself (respondents and non-respondents), from the geographical context, and from auxiliary variables in the population (Groves 2006).

Based on the described operational definitions of bias, research activities⁵ are organized around four ways of bias detection and estimation in ESS. These are: (1) post-stratification weighting (Vehovar 2007); (2) comparison of cooperative with reluctant respondents (Billiet et al. 2007a, b); (3) using information from observable data (Stoop et al. 2010, pp. 243–252); and (4) obtaining information from additional surveys among non-respondents (Matsuo et al. 2010b).

Each of these approaches has its own advantages and disadvantages (Billiet et al. 2009): (1) post-stratification approach using population data/statistics is useful for adjustment but this information is limited to variables such as age, gender, and education. However, this information is not always available and reliable in all countries and the relationship with target variables is rather weak; (2) many variables are available for reluctant respondents but non-response can be only adjusted for refusals (subgroup of non-respondents) and not for all non-response; moreover, it is questionable how far the reluctant respondents are representative for all refusers; (3) observable data (dwelling and neighborhood variables) is available for all sample units including both respondents and non-respondents but the information is limited to a small number of variables (type of house, physical condition of house, presence of vandalism, and presence of litter); missing values in these variables are high, the measurement is not very reliable, and the relationship with target variables is also weak; (4) a survey among non-respondents some months after the fieldwork of main ESS round contains small number of (relevant) questions (16) but is only available for four countries. The last approach applied propensity score stratification method (10 classes)

on the basis of the information of co-operative respondents and non-respondents but the weighting was not very effective (ESSI workshop 2011 review).

10.2.3 Increasing Response or Reducing Bias?

Survey response rates are generally decreasing (de Heer 1999), and non-response rates increase including refusals and non-contacted sampled persons. This may affect bias in the estimates. However, response rates are insufficient indicators of survey quality since what matters most is the extent to which respondents differ from non-respondents on crucial (target) variables in a survey. Against this background, the 'Risq' project⁶ has developed the R-indicator or representativeness indicator. This indicator allows for comparing the quality of surveys between populations and over time. It also constitutes a valuable tool for monitoring the quality of the fieldwork in accordance with the concept of 'responsive design' (Groves and Herringa 2006).

The unique contribution of the R-indicator is situated in the sample-based models for individual response propensities.⁷ R-indicators are random variables based on the standard deviation of estimated response propensities, and the modeling of response propensities has important implications. This means that the sample size and availability of auxiliary variables play an important role in the estimation of response propensities. When survey responses are representative, all response propensities have equal value, whereas, when the survey response is not representative, response propensities have large variations.

The kind of variables used for the R-indicator can differ between surveys in terms of focus and extent. Use is usually made of auxiliary information like the socio-demographic characteristics of the sample unit, sample unit dwelling and

⁶ This project is financed by the EU 7th Framework Programme and involving a number of European institutions.
⁷ Schouten et al. (2009) discusses the mathematical features of R-indicators. Risq project (<http://www.risq-project.eu/indicators.html>)

neighborhood characteristics, or the results of every attempt to contact the sample unit. These are available from sample frame data, censuses, or administrative data, to contact files based on interviewer observation. Needless to say, such variables must be available for all sample units outside the survey questionnaire and are constant during the fieldwork period. Considered variables are: age, gender, ethnicity, degree of urbanization, type of household, region, province, existence of listed phone number at household level, average house value at zip code level, welfare status (receiving disability or social allowance), employment status (Schouten et al. 2009) as well as treatment variables (or so-called key process variables) such as number, timing, mode of contact attempts, and interviewer skills (Loosveldt and Beullens 2009).

10.2.4 Evaluation of the Obtained Sample in ESS: Illustrations

This section presents three applications of the evaluation of the obtained sample in a cross-national survey. In the first application, different response and non-contact rates are compared between four rounds of ESS. This is a typical example of output evaluation. The second application deals with one aspect of the response enhancement strategy, namely, the refusal conversion process, which is an example of process evaluation. The third application is an example of output evaluation assessing the level of non-response bias by comparing cooperative respondents with reluctant respondents. The evaluation of the obtained samples in ESS took place in the context of general ESS research activities and the aforementioned *Joint Research Activities* (JRA) on non-response bias (Billiet et al. 2009).

10.2.4.1 Illustration 1: Assessing Response and Non-Contact Rates

Response rates are most usefully assessed with regard to a benchmark, usually a minimum response rate target. The ESS minimum response rate target is 70% while the maximum non-contact rate is 3%. These targets are not always

reached. In Round 1, the number of countries reaching (or almost reaching) the 70% minimum response rate target is small (four countries) yet significant cross-country variation exists as response rates range from 32.5% (CH) to 79.5% (GR). The mean response rate is 60.3% and the dispersion (standard deviation) is 12.6 points. Two rounds later, the average Round 3 response rate for 24 countries increases to 62.5% with a standard deviation of 7.9 points, whereas a decline in the average response rate among 30 countries (0.6% points) and an increase in the standard deviation of 2.2 points are observed between Rounds 3 and 4. The spread between the lowest (42%) and the highest (81.0%) response rates is the second highest (38.3% points) of the four rounds. Target non-contact rates, however, are not always reached either. It should be noted that low non-contact rates (maximum 3%) is more easily achieved than high response rates (minimum 70%). The number of countries reaching the maximum 3% non-contact rate increases by each rounds (from 5 in Round 1 to 13 in Round 4). In Round 1, the mean non-contact rate is 5.4% and the dispersion (standard deviation) is 3.5 points. In Round 3, the average Round 3 non-contact rate for 24 countries decreases to 4.6% with a standard deviation of 3.1 points. In Round 4, there is an increase in the average non-contact rate among 30 countries (0.6% points) and an increase in the standard deviation of 1.0 points (Matsuo et al. 2010a, p. 55) (Table 10.1).

10.2.4.2 Illustration 2: Evaluating Response Enhancement Strategy—Refusal Conversion Process

As explained before, refusal conversion consists of reproaching initially reluctant respondents in order to persuade them to reconsider participating in the survey. Refusal conversion can boost the response rate since it can be assumed that many initial refusals are 'soft' refusals instigated by particular circumstances (e.g., bad timing, inappropriate interviewer tailoring technique, etc.) that may not apply at a later occasion.

In ESS, substantial cross-country variation existed with respect to both the proportion of

⁵ These research activities are ESSI (infrastructure) program funded project of 'Joint Research Activities (JRA2): improving representativeness of the samples' implemented in 2006–2011.

Table 10.1 ESS round 1 (2002)–round 4 (2008)

Response rates	Round 1	Round 2	Round 3 ^a	Round 4
Mean%	60.3	62.6	62.5	61.9
STD	12.6	9.6	7.9	10.1
Lowest%	32.5	43.6	46.0	42.7
Highest%	79.5	78.8	73.2	81.0
Range	47	35.2	27.2	38.3
# more than 70%	4	6	4	7
Non-contact rates				
Mean%	5.4	5.4	4.6	5.2
STD	3.5	3.0	3.1	4.1
Lowest%	0.8	0.9	0.8	0.0
Highest%	14.7	10.9	13.1	14.1
Range	13.9	10.0	12.3	14.1
# less than 3%	5	7	10	13
N (countries)	20	23	24	30

Reproduction of table based on Matsuo et al. 2010a

Average response and non-contact rates (% of total eligible sample sizes) and characteristics of the response rate distributions at country level (Figures based on the analysis of the call record data and/or National Technical Summary (NTS) (see Matsuo et al. 2010a, p. 55))

^a Country with largest response rates is missing in Round 3

initial refusers contacted for refusal conversion and the refusal conversion success rate. Table 10.2 shows the effect of refusal conversion presenting figures on the number and proportion of initial refusers, the proportion reapprached, the number of converted respondents, the success rate of the conversion attempts, and the effect on final response rates. Both the share of initial refusers reapprached for refusal conversion and the refusal conversion success rate depend to a large extent on the nature of national refusal conversion administrative and fieldwork specifications on the identification, selection, and contacting of reluctant respondents. Examples are given from the most recent round (Round 4) as it has the highest number of participating countries implementing refusal conversion attempts (see Table 10.2).

Substantial cross-country variation is first of all observed with regard to the coverage of refusal conversion activities (percentage of initial refusers that are reapprached). A large

proportion (>80%) of initial refusers was reapprached in some countries (e.g. NL and CH) while very limited refusal conversion efforts (<10%) were made in the other 10 countries.⁸ For example in The Netherlands, the initial refusal rate is high (more than 50%) and approximately 90% of the initial refusers are called on again. Hardly any initial refusers are reapprached in the aforementioned ten countries with an initial refusal rate ranging between 4.6 and 37.9%.

The refusal conversion success rate also differs substantially between countries. This is the share of reapprached initial refusals which were successfully converted. However, one must interpret this carefully. Some countries target their refusal conversion activities to a restricted number of initial refusers, whereas other countries reapprach substantial numbers of initial refusers (e.g. more than 50% of initial refusers).

⁸ BG, CY, CZ, DK, GR, LV, LT, PT, RO and UK.

Table 10.2 Outcome of refusal conversion attempts based on the analysis of contact files, Round 4

Country	Initial refusers (N)	Initial refusers (%)	Reapprached (%)	Converted respondents (N)	Success rate (%)	Effect on response rate ^a
BE	844	28.3	67.3	70	12.3	2.3
BG	493	16.6	4.5	17	77.3	0.6
CH	1535	41.2	86.5	291	21.9	7.8
CY	151	10.0	4.6	2	28.6	0.1
CZ	778	26.8	2.4	0	0.00	0.0
DE	2709	42.0	55.9	611	40.4	9.5
DK	986	33.1	0.6	5	83.3	0.2
ES	820	21.2	58.2	198	41.5	5.1
FI	750	23.4	59.6	80	17.9	2.5
FR	1614	38.8	56.8	106	11.6	2.5
GB	1197	27.8	58.3	146	20.9	3.4
GR	502	18.0	0.2	1	(100) ^b	0.0
HU	758	30.1	50.7	67	17.4	2.7
HR	949	29.4	41.1	170	43.6	5.3
IE	963	26.8	41.4	93	23.3	2.6
IL	281	8.7	34.5	27	27.8	0.8
LT	1347	37.9	0.0	0	0.00	0.0
LV	406	11.6	2.0	4	50.0	0.1
NL	1882	52.7	91.2	435	25.3	12.2
NO	892	34.8	54.6	101	20.7	3.9
PL	479	21.0	56.8	69	25.4	3.0
PT	627	20.1	0.0	0	0.00	0.0
RO	144	4.6	1.4	0	0.00	0.0
RU	986	26.4	36.0	53	14.9	1.4
SE	822	28.0	46.4	79	20.7	2.7
SI	553	25.3	37.3	77	37.4	3.5
SK	481	19.3	57.6	161	58.1	6.5
TR	719	18.9	19.3	13	9.4	0.3
UA	753	25.1	3.6	8	29.6	0.3

Estonia was excluded from the analysis due to the lack of information about refusal conversion process. On the basis of Matsuo et al. (2010a), this table takes a similar format as reported in Stoop et al. (2010), p. 172

^a Same calculation applied expressed as percentage point differences relative to initial refusal rate

^b Note the very small proportion of reapprached

Countries reappraching substantial number of initial refusers combined with conversion rate that exceeded 2% are observed in 8 countries.⁹

The final effect on response rates is shown in the last column of Table 10.2 as a percentage

increase of response rate with and without refusal conversion attempts. The impact of refusal conversion attempts on final response rate also differs substantially between countries. For some countries, the impact of refusal conversion on final response is major. Substantial changes in response rates (3+ or more) are observed in ten

⁹ CH, DE, ES, GB, NL, NO, PL and SK.

countries.¹⁰ Most of these ten countries, except for Poland and Slovenia, obtained more than 100 reluctant respondents. In six other countries,¹¹ changes of 2–3% points are observed. For the remaining 13 countries, the impact is minor as changes of <1% are observed.

10.2.4.3 Illustration 3: Comparing Cooperative and Reluctant Respondents

One way of assessing non-response bias, consists of taking a closer look at the category of converted initial refusers or "reluctant respondents". Like a comparison between respondents and non-respondents, a comparison between "reluctant respondents" and "cooperative respondents" (respondents who agree to survey cooperation without refusal conversion activities) can provide an insight into the presence and extent of non-response bias.

The case study (Stoop et al. 2010, pp. 234–238) presented here pertains to ESS Round 2 and focuses on countries with comparatively high numbers (>100) of reluctant respondents: Switzerland, Germany, Estonia, The Netherlands, and the Slovak Republic.¹² Reluctant respondents were compared with cooperative respondents for a number of background, attitudinal, and media use variables and for responses to a number of survey questions that were marked by a rather large average absolute standardized bias according to the post-stratification approach (Vehovar 2007).

A multivariate logistic regression model was used to assess the differences between the two types of respondents (reluctant vs. cooperative) (Table 10.3). The dependent variable was the odds ratio between the two types of respondents. Target persons from larger Swiss households were initially more reluctant to participate but more often persuaded to cooperate at the refusal conversion stage. In Germany, converted refusals were significantly more likely to be women,

to be older, to live in large cities, to engage in Internet surfing, and to have a history of unemployment, and less likely to participate in politics. In Estonia, reluctant respondents were more likely to be women and to have a paid job, while living in a village and having ever been unemployed were associated with a greater likelihood of being a cooperative respondent. In Slovakia, compared to cooperative respondents, converted refusals were more likely to be old, to have a medium level of education (upper and post-secondary education), to be more religious, and to feel comfortable with their household income.

On the other hand, they were less likely to have ever had a job and to feel safe walking in the neighborhood after dark. In The Netherlands, the likelihood of being a converted refusal increased when the respondent was female, had an average upper and post-secondary level of education, watched more television, and surfed more frequently on the Internet. Converted refusals were also more likely to have "ever had a job", to feel healthier, to see immigrants as a threat, and to be dissatisfied with their own life. However, unexpectedly given the literature, respondents who trusted political institutions, obeyed the law, participated in political organizations, and felt socially integrated were somewhat more likely to be converted refusals (Stoop et al. 2010, pp. 234–38).

10.3 Measurement Error

The conception of survey measurement error and the way to study it has been changed over time. After the early "split ballot" approach, the study of measurement error changed in the direction of mathematical modeling (Rossi et al. 1983, p. 17). Until the 1990s the conception of survey errors was statistical rather than (cognitive) psychological in character. The statistical models paid more attention to the variance component of the error as is the case in the Groves' (1989) standard work on survey errors and survey costs. Since the CASM¹³ movement

¹³ CASM: cognitive aspects of survey methodology.

¹⁰ CH, DE, ES, GB, HR, NL, NO, PL, SI and SK.

¹¹ BE, FI, FR, HU, IE and SE.

¹² Slovenia was excluded from the analysis due to data quality issue although the number of reluctant respondents was high.

Table 10.3 Statistical significant (<0.05) Logistic regression estimates (odds ratios) for reluctant versus cooperative respondents, ESS 2

	Switzerland	Germany	Estonia	Slovakia	The Netherlands
<i>Background variables</i>					
Male (= Yes)		0.797 ^c	0.706 ^b		0.614 ^a
Age		1.019 ^a		1.019 ^a	
Number of household members	1.177 ^b				
Urbanicity					
Countryside and village	0.634		0.677 ^b		
Town and small city	0.816 ^b		1.293 ^c		
Large cities and suburbs	1.312 ^a		1.143 ^a		
Level of education					
Lower and lower secondary				0.682	0.817
Upper and post-secondary				1.758 ^b	1.243 ^b
Second stage of tertiary				0.834	0.984
Labour market status					
Paid job (1 = yes)			1.863 ^a		
Ever job (1 = yes)				0.425 ^c	1.694 ^c
Ever unemployed (1 = yes)		1.357 ^c	0.584 ^c		0.589 ^c
Good health (1 = yes)					1.296 ^c
Comfortable income (1 = yes)				1.887 ^c	
Religious involvement (0–10)				1.534 ^a	
<i>Attitudes</i>					
Perceived ethnic threat (0–10)					1.136 ^a
Trust political inst. (0–10)					1.101 ^c
Political participation (0–10)		0.911 ^c			1.096 ^c
Civil obedience (0–10)					1.067 ^c
<i>Satisfaction and integration</i>					
Attitudes					
Social isolation (0–10)					0.862 ^a
Safe after dark (= yes)				0.464 ^b	1.097 ^c
Media use					
TV watching (minutes/day)					
Internet use (no access to every day = 0–7)		1.065 ^b			1.001 ^a

^a $p < 0.001$

^b $p < 0.01$

^c $p < 0.05$

Only statistically significant parameters are reported ($p < 0.05$) for all variables, except for categorical variables where all parameters are reported when one of the categories has a significant effect on the response variable

(mid 1980s) researchers have increasingly traced measurement errors to the psychological processes of the respondent. The new developed cognitive models do not so much contradict the earlier models of error but supplement them and

pay much more attention to the bias component. Central to measurement error is the idea that there is no perfect relationship between the observed response and the underlying unobserved

variable that serves to generate the response (Borinestad 1983, p. 70). The obtained measure is a (linear) combination of the true value, the effect of the method used, and random error. A very general distinction in the literature of measurement error is between *reducers* and *measures* (Groves 1989, p. 4). Applied to the survey instrument (the questionnaire), the *reducers* try to reduce error as much as possible by applying rules for appropriate phrasing of survey items. Examples of this approach are in the classical volumes on the art of asking questions and questionnaire design of Payne (1951), Sudman and Bradburn (1974, 1982), Converse and Presser (1986), Foddy (1993), and more recently in Saris and Gallhofer (2007, pp. 81–153). This work applied to questions in standardized questionnaires and guidelines are offered for different kinds of questions. Many of the suggestions and rules are based on research findings of the measures on the effect of variations in question wording. Reducers and measures are thus not necessarily different scholars, but different ways of dealing with survey questions and responses. The measures try to measure measurement error in order to understand it, to identify the causes of error, or to obtain estimates (e.g. correlations) that are adjusted for measurement error.

Among the measures one can distinguish several approaches that are discussed in the remaining part of this chapter. It starts with the findings, advantages, and disadvantages of measuring the effects of the wording and order of questions and response scales in the so-called "split ballot designs". The contribution of the CASM movement to the understanding of some of the major findings of split ballots is then discussed. After a critical appraisal of the split ballot approach we move to the modeling approach in which much attention is paid to the "multitrait-multimethod" design used to assess the reliability and validity of questions, and to the modeling of response styles.¹⁴

¹⁴ The challenge of the cross-cultural equivalence of measurement models is not considered here. For this, we can refer to "Cross Cultural Analysis: Methods and Applications" (Davidov et al. 2011).

10.3.1 The Split Ballot Approach to Measurement Error

In the classical field experiments of the measurers, the so-called "split ballots", questions are formulated in different ways (Rugg 1941; Cantril 1944). The variants are administered to equivalent random samples of respondents and the response distributions are compared. Sometimes panel data is used, i.e., the same respondents are interviewed on two or more occasions with different question forms. The experiments are mostly carried out in the context of general social surveys in order to combine the testing of causal hypotheses through experimental manipulation (internal validity) with the ability to generalize to the national population (external validity) (Schuman and Presser 1981).¹⁵ Not all findings about what can go wrong with questions come from split ballots. Some researchers like Belson (1981) systematically investigated the meaning of the words used in standardized questionnaire by asking respondents how they understood the questions. This method was adopted in later cognitive interviewing (Forsyth and Lessler 1991).

The most debated response effects deal with the following formal characteristics of questions: the length of questions, open and closed questions, yes/no-questions versus choice questions, contrasting questions, the use of arguments, questions with or without the middle-alternative, the "no opinion" option and the use of filters, the number of the answer categories, the order of the alternatives, the overall position of the question in the questionnaire, and the sequence of the questions (Molenaar 1986; Kalton and Schuman 1982). It seems that nearly every type of variation in question form and wording has an effect on response distributions.

10.3.1.1 Major Findings from Split Ballot Experiments

Let us start with a general view on the major findings from split ballot field experiments with

¹⁵ A revised edition appeared in 1996, Thousand Oaks Ca.: Sage Pub. Inc.

variations in the wording and order of questions and response categories. For a short overview see the introductory chapter of Foddy (1993, pp. 2–10).

(1) Slight changes in the tone of wording produce major changes in the distribution of responses. This is in general the case with opinion questions, but answers to questions about the respondents' characteristics or about past behavior are also not completely protected from this (Foddy 1993, pp. 4–7). A well-known example of the effect of small variation is the difference in acceptance of policy measure depending on the verb "forbid" or not "allow" in a "yes or no" format (Schuman and Presser 1981, pp. 276–286). This kind of response effect finds a favorite application in the context of referenda with pairs of verbs as 'revoke/maintain', 'revoke/uphold', or 'facilitate/prevent' (Beckers and Billiet 2010). Hollenman (2000) offered in her doctoral study a plausible (cognitive) explanation of this kind of response effects: there is a mental latent continuous scale behind the yes/no dichotomy and the respondent is moving on this scale depending on the terms used.

(2) The emotional and social pressure on the respondents to choose an answer that is in accordance with the suggested direction of persuasive questions can emerge from different formal attributes of a question. One-sided arguments, suggestive words, or unbalanced response categories are likely to change the response distribution in the direction of the persuasion (Schuman and Presser 1981, pp. 179–199; Molenaar 1986, p. 166).

(3) Respondents' answers to questions are sometimes affected by question format per se. Open questions without pre-listed response alternatives produce quite different results from closed questions, even when an exactly identical question text is used (Foddy 1993, pp. 7–8). Not seldom, less than 50% of the responses to offered categories in a closed question are mentioned

on the closed form (Schuman and Presser 1979, p. 696). In a closed format, response alternatives may remind respondents of opinions that they may otherwise not contain in a free-recall task. Another explanation points to the possibility that the set of response categories affects the interpretation of the question's content. Presenting respondents a list of activities reduces the likelihood that respondents report activities that are not on it. Respondents may assume that the researcher is not interested in information that comes spontaneously in their mind but that is not included in the list (Schwarz and Hippler 1991, pp. 43–45).

(4) Responses to closed questions with the request to indicate for each listed option in a long list whether it applied or not, differs considerably from closed questions with the request to indicate from a list all options that apply. Response distributions over time (or between countries) are not comparable when the format has been changed. We found a striking example of this response effect in the 2008 *European Value Study* in Belgium. The shift in format happened in the field organization at the moment that the paper version of the questionnaire was transformed into an electronic version in view of CAPI. This shift in format resulted in serious underestimation of responses to questions containing lists of characteristics of a job, preferred qualities of children, and groups in the population the respondents do not want as neighbors. The results are no longer comparable with previous waves and with other countries that participated in EYS 2008 because of this shift in format (Billiet 2011).

(5) Even small changes in the number and wording of the response categories had serious effects on the response distributions (Schuman and Presser 1981, pp. 161–178). This is for example the case when a logical (and realistic) middle alternative is omitted (see later Table 10.2). This sensitivity of

response distributions for changes in the wording and form of response scales makes it indispensable to screen always the response categories when response distributions of (seemingly) identical questions are compared between groups over time.

- (6) Response scales can provide not intended information, and may have an effect on the respondents' choice of a response category. If questions about the frequency of engagement in activities (by themselves or others) are asked (*how often did you...*) and a response scale of specific length is offered, then the respondents seem to use the range of the numeric values given on the rating scale when they select a (frequency) category. (Schwarz and Hippler 1991). This response effect was found in split ballot experiments with high and low range response scales, for example about the daily time of watching television. Apart from the meaning of "watching" television, the range of the scales informs the respondents about the likely answer. This results in much higher frequencies when a high range is offered. When respondents are asked to check the appropriate response category they may infer from the range of response alternatives that a particular form of behavior is typical or socially desirable (Schwarz et al. 1988; Schwarz and Hippler 1991, p. 48). Respondents assume probably that the researcher offers a fair scale and that the middle of the scale reflects the behavior of modal people while the extremes represent the "outliers". Verbal anchors of numeric response scales also may carry unintended information and affect the respondents' choice (O'Muircheartaigh et al. 1995).

- (7) Respondents often answer questions even when it appears that they know very little about the topic (Foddy 1993, pp. 8-9). It has been found that virtually all the respondents chose substantive response options when the "Don't know" (DK) option is not offered, but a substantive

number of respondents (up to 1/4) select that option when it is offered. This finding stimulated a lengthiness discussion on 'non-opinions' or 'non-attitudes'. Does a large portion of the respondents really have weak or even no opinions at all on most issues (Converse 1964)? Is it an indication of privacy protection? Or is the un-stability related to the fact that respondents have several opinions depending on the considerations they have in mind at the moment of their choice (Zaller 1992)? This response effect and its explanations relate to the finding that answers to attitude questions are very unstable over time.

- (8) Respondents' reported opinions, beliefs, choices, habits, and interest often seem to be very unstable. A study of the reliability of survey measurement estimated the test-retest reliability of 41 attitude items from the *US National Elections Study* panels (1950-1970) to be 0.55 on average (Smith 1984). Abstract attitude questions are less stable than more concrete ones, and the lower educated and respondents with weak attitudes are most unstable. Latent multiple indicator variables are more stable than single items (Billiet et al. 2004). Several explanations are offered for the observed instability of attitude measurements. As was mentioned, Converse suggested that many respondents "simply do not have meaningful beliefs even on issues that have formed the basis for intense political controversy among the political elite for substantial periods of time" (Converse 1964, p. 245). Others contend that instrument unreliability and measurement error, rather than non-attitudes, were the cause of low correlations (Alwin 1992). It is argued that the underlying "true" attitudes are more stable (see Smith 1988 for an overview), and that the observed changes in responses to opinion questions are due to changing considerations (Zaller 1992, 42-51).

- (9) Answers to earlier questions can affect respondents' answers to later questions.

These are so-called "context effects". The context of a question, that is the preceding question, and response scale, plays a crucial role in determining the meaning of that question. Preceding questions and responses can modify the meaning of the later questions, and responses to later questions(s) are then consistent or in contrast with prior questions. A well-known example of a contrast effect is the change in acceptance of abortion in a more general context (less acceptable situation) after a very specific question dealing with a more acceptable situation. When the item with the acceptable reason for abortion is asked first, respondents may assume that the specific (acceptable) argument is not included in the subsequent general question and it then becomes easier to oppose the general item (Schuman and Presser 1981, pp. 36-39; Billiet et al. 1992, pp. 134-135).¹⁶ The scope of the term "abortion" has changed because the most acceptable reason is mentally left out (mental subtraction). Previous questions not only evoke a change in the meaning of later questions, but they can also increase the saliency of a topic (Schuman and Presser 1981, pp. 44-47).

- (10) Changing the order in which response options are presented affects respondents' answers. Primacy (choosing first option) or recency (choosing last option) effects appear depending on the mode of administration (face-to-face, mail, telephone) (McClendon 1991) and the use of response cards where visual display of the categories is offered (Schuman et al. 1996, pp. 138-142). The alternatives to a closed question are mostly asked in the same order for all respondents and the effect is only detected when for one or another reason, the order has been changed in comparable samples. Researchers

¹⁶ The standard interview situation, where it is assumed that the interviewer does not ask the same question twice (the 'given new context' rule) is a favorable context for this kind of context effects.

then realize that the choices made by the respondents are artifacts of response order. It is therefore suggested to randomize the order of response categories in order to have at least an idea about the size of the effect.

- (11) The cultural context in which a question is presented often has an impact on the way respondents interpret and answer questions. This context might be mediated by the main subject of the questionnaire, by the presence of the interviewer, and even by the ongoing debates in the media. Independent of the question context, different meanings may be differentially accessible to different respondents because of the frequency with which they employ them in daily life (Sudman et al. 1996, p.61).¹⁷

10.3.1.2 Initial Understanding Effects of Question Wording Effects

In an attempt to understand the effects of question form on response marginal, Schuman and Presser (1981, pp. 298-301) suggested two concepts that are each applied to several classes of response effects. These are *question constraint* and *response persuasion*.

Question constraint applied to several of the findings that are mentioned before: omitting a don't know or a middle alternative (see Table 10.2), changing the number response categories, open versus closed questions, etc. In the context of survey interviewing with standardized questionnaires, respondents seem to accept the "rules of the game" as they are conveyed by the form of the questions, and they earnestly try to work within that framework once they have agreed to be interviewed. The

¹⁷ In a survey in the 1980s, we have observed that some respondents offered numbers between 20 and 30 in response to the question "How many children do you have?". These respondents were school teachers who interpreted the question to refer to the children in their classes, the meaning that was most accessible in their memories (cited in Bradburn, 1992, p. 317). That question was in the last part on the household characteristics of teachers in a self-administered questionnaire about their attitude toward mixed (qua gender) classes.

Table 10.4 Response distributions of two questions on in-group favoritism in two forms (split ballot: with and without a middle alternative) in two random samples of the Flemish population in 1995 (in percentages)

Form A (no MA included)		Form B (explicit MA included)	
"Suppose that there are two employees, one is Flemish, the other belongs to an ethnic minority. Both are equal on all characteristics. When one of them has to be dismissed because the company's financial situation, who should be dismissed according to you, the Flemish or the employee belonging to an ethnic minority?"		"Suppose that there are two employees, one is Flemish, the other belongs to an ethnic minority. Both are equal on all characteristics. When one of them has to be dismissed because the company's financial situation, who should be dismissed according to you, the Flemish or the employee belonging to an ethnic minority? Or should this made no difference?"	
The employee of ethnic minority	45.1	The employee of ethnic minority	25.9
The Flemish employee	7.4	The Flemish employee	6.4
Should not made any difference*	39.5	Should not made any difference	65.1
No opinion*	7.9	No opinion*	2.6
N (100%)	532	N (100%)	455

$\chi^2 = 70.087$; $df = 3$; $p = 0.001$		"Suppose that there are two employees, one is Flemish, the other belongs to an ethnic minority. Both are equal on all characteristics. When one of them has to be dismissed because the company's financial situation, who should be dismissed according to you, the Flemish or the employee belonging to an ethnic minority?"	
"Suppose that there are two employees, one is Flemish, the other belongs to an ethnic minority. Both are equal on all characteristics. When one of them has to be dismissed because the company's financial situation, who should be dismissed according to you, the Flemish or the employee belonging to an ethnic minority?"		"Suppose that there are two employees, one is Flemish, the other belongs to an ethnic minority. Both are equal on all characteristics. When one of them has to be dismissed because the company's financial situation, who should be dismissed according to you, the Flemish or the employee belonging to an ethnic minority? Or should this made no difference?"	
The employee of ethnic minority	4.0	The employee of ethnic minority	0.8
The Flemish employee	47.1	The Flemish employee	31.7
Should not made any difference*	43.0	Should not made any difference	65.5
No opinion*	6.0	No opinion	2.0
N (100%)	532	N (100%)	455

$\chi^2 = 57.195$; $df = 3$; $p = 0.001$

* Not included in the question text but spontaneously answered

constraint imposed by the question is very powerful, and the researchers' decision to omit an alternative, or to modify the wording, has considerable influence in shaping the resulting response distributions (Schuman 1984, p. 34). At first glance survey research practice may find a rule is this understanding of response effects: remove the constraint in questions. It is however not that simple as was obvious in the discussion of Table 10.4. There is even not one simple rule for the DK filter: If one wants to measure clearly informed opinions, then one may include an implicit DK filter, but if one wants to obtain the responses to a set of indicators (items) that is intended to measure an underlying disposition (or a latent attitude variable), then the DK filter is not recommended (see Schuman and Presser 1981, pp. 147–160).

Response persuasion deals with the persuasiveness of the question text or with response categories or scale. Emotional and social pressure on the respondents to choose an answer that is in accordance with the suggested direction can emerge from different attributes of a question: questions with one-sided arguments; clauses and emotionally loaded terms in a question; reference to opinions of others; answer categories that express positive judgments; unbalanced questions (i.e. no explicit mention of a negative alternative of a statement). A general rule for survey practitioners is more evident here: never deliberately load questions. But there are even exceptions in this case. It is sometimes suggested to use leading questions in case of expected underreporting of social undesirable responses to factual questions, albeit with care

(Sudman and Bradburn 1982). Experimental variation of (counter-) arguments in repeated measurements (in web surveys) may be used for measuring the stability and crystallization of opinions, and should be distinguished from deception (see Sniderman and Grob 1996).

Several response effects that were listed but are not covered by these two concepts received full attention in the CASM movement. This is the case of *unintended information* provided by the researcher through the response scales. Respondents use the information in the response scales to interpret the meaning of a question, and to find out what kind of information the interviewer wants, or what is the 'correct' answer.

Context effects became in the 1980s a favorite subject for theory-based cognitive experiments with survey questions in which interpretations of these effects were tested (Schwarz and Sudman 1992, pp. 28–44; Sudman et al. 1960). These are considered as shifts of the respondents' *frame of reference*. The context plays a crucial role in determining the meaning of a question. Preceding questions can modify the scope of subsequent questions (e.g. the abortion questions). Previous questions may also make some responses to later questions more accessible (by priming), or may increase the saliency of questions (Tourangeau 1992).

10.3.1.3 Social Cognition and Question Wording Effects

Schuman and Presser reacted initially very skeptically to the recommendation that research on question wording should draw its theory from cognitive and social psychology. They suggested that a substantial contribution to survey research would be made only if those disciplines become directly involved in studying survey data, and formulate problems that arise directly from the data, methods, and ideas of survey researchers (Schuman and Presser 1981, pp. 313–314). A number of social and cognitive psychologists accepted the challenge (Jabine et al. 1984).

Considerable progress in understanding the cognitive and communicative processes underlying survey responses has been made over the past 25 years (Schwarz et al. 2008). Scholars in social

cognition tried to "find the substantive meat in what others may well regard as methodological poison" (Tourangeau and Rasinski 1988, p. 301). The standardized question situations of surveys offer excellent databases to psychologists who studied the cognitive determinants of responses to questions. In such a restricted conversational interaction that excludes negotiation about the intended meaning of questions, the respondents rely to a large degree on previous questions and answers (the context), on the formal characteristics of the target question, and—in case of face-to-face interviewing—on the instructions and cues provided by the interviewer as sources of information (Strack et al. 1991, pp. 112–113).

Several theoretical models of the cognitive and communicative processes underlying responses to survey questions have been proposed in collaborative research programs of (cognitive) psychologists and survey methodologists (Biderman 1980; Hippler et al. 1987; Schwarz and Sudman 1992). Several volumes are published in which the results of question wording experiments are theoretically grounded in the social cognition tradition. "Thinking about Answers" (Sudman et al. 1996) and "The Psychology of Survey Response" (Tourangeau et al. 2000) offer comprehensive syntheses of the contribution of (cognitive) psychology to the understanding of survey response. This work mainly relates to questionnaires in face-to-face or telephone interviews. More recent developments apply to web surveys (Couper 2001; Dillman 2001; Heerwegh 2003), and even surveys by mobile phones (Lynn and Kaminska 2011).

According to the model of information processing in a survey interview (Strack and Martin 1987), respondents need first (1) to *understand* the question and determine which information they are asked to provide. Further, they need to (2) *recall* the relevant information from memory. They need to (3) form a *judgment* on the spot based on the information that comes to mind at that time and that is considered relevant and useful for answering the question. In case of factual questions on events or on past behavior,

respondents need to retrieve relevant episodes. Once a judgment is formed or relevant episode is retrieved, they rarely report it in their own words but they need (4) to *format* it in the response alternatives offered. Respondents may however hesitate to (5) communicate (*editing*) their 'internal' response because of social desirability and self-presentation (Schwarz et al. 2008, pp. 18–19). The response effects that we discussed are related to each of these five tasks and can be explained by several cognitive psychological processes in interaction with the instrument (questionnaire), and the communication context of the interview (e.g. priming, information accessibility, the inclusion/exclusion of information.). "Satisficing" occurs when respondents "shortcut the cognitive processes necessary for generating optimal answers" (Krosnick et al. 1996, pp. 29–31). When a respondent works through each of the five mental steps with a maximum level of effort, they are optimizing their answers.

10.3.1.4 Response Effects as Useful Data: An Appealing Example

A pair of questions on in-group favoritism was included in a survey on political knowledge organized by the *Institute for Political Opinion Research* (ISPO), in a random sample of the general population in Flanders in 1996 ($N = 987$). It was decided to split up the sample into two random parts according to the absence (Form A) or the presence (Form B) of the middle alternative (MA) in questions and response cards. The response distributions of the discrimination questions in the two subsamples are in Table 10.4. Comparison of the response distributions in the two forms suggests that not all respondents who want to dismiss (45.1%) the employer of an ethnic minority or offer promotion to the Flemish employee (47.1%) have a fixed opinion. We are not sure since these are split samples and not repeated measurement, but this conclusion is very likely. The percentage of respondents who are ready to give preferential treatment to a person of their own ethnic group is 15–20% points lower when the option "should not make any difference" is offered.

measures of the same underlying latent variable. It is possible to cross the forms over the two subsamples. Each sample contains then the A and the B forms, but it is important to have the B form with MA second in the two samples, in order to avoid priming of the middle alternative and creating that way a consistency effect.

10.3.1.5 Critical Reflections on the Split Ballot Approach

The split ballot experiment is a strong design for the assessment of a question wording effect. However, this approach had some shortcomings. A finding holds only for a specific question involved in the experiment. To be able to generalize a particular wording effect to other questions, many similar experiments on the same wording variable would be required, the results of which could then be summarized and tested for significance. The number of investigations is not large enough to warrant general statements about the effects of particular wordings (Molenaar 1986, p. 161). Moreover, it is not always clear what the most valid wording is. We have seen in our example of in-group favoritism that it depends on the research question in a substantive research context.

As an alternative, Molenaar (1986) worked out a *non-experimental design* that consists of a meta-analysis of hundreds of questions and distributions according to formal characteristics of questions. He also paid serious attention to linguistic characteristics of questions. One should realize that the critical view of Molenaar on the split ballot experiments apply to these in the period before the fertilization of these experiments by the insights from cognitive psychologists. The findings about particular question wording effects became more general in the 1980s by enriching these by their theoretical insights in cognitive processes, and by setting up theory-driven experiments (see Sudman et al. 1996).

An important development that changed the nature of split ballot experiments, and that overcomes the ad hoc and a theoretical character of the early split ballots was the introduction of computer-assisted interviewing and the possibility of web surveys. Sniderman and Grob

(1996) argued that the introduction of computer-assisted interviewing (in whatever mode) has changed the opportunities of the classical split ballots considerably. A new wave of experimentation in survey research was made possible by the introduction of computer-assisted interviewing and surveys via the Web. This new wave of experimentation is characterized by a double shift: from fixed and predominantly methodological to variable forms and predominantly substantive focus (Sniderman and Grob 1996, p. 380). Randomization of samples and several versions of questionnaires became much easier (Sniderman and Grob 1996, p. 379). The computer-assisted experimental design has been applied in large-scale general population samples. Questions as "how stable are these opinions for which subgroups in a population?" or "how likely are these to change when some considerations are primed?" can now be answered. The effects of several relevant considerations (and combinations) can be studied by introducing variation in the framing of the choice presented to respondents, and variation in the context of questions (Sniderman and Grob 1996, p. 393). Researchers who apply this strategy should be very open-eyed not to cross the fine line between experimentation and manipulation, and to avoid deception of respondents.

10.3.2 Modeling Measurement Error

The initial models used for quality assessment of measurement instruments are the *Quasi Simplex* (QS) approach and the *Repeated Multinomial Model* (RMM) approach (Saris and Andrews 1991, pp. 575–548). The QS approach originated from Heise's (1969) idea that test-retest studies could be the most appropriate approach for evaluating survey questions in panel research in which traits were repeatedly measured with the same method, and was used by Alwin (1989) within the framework of error sources in survey data. The QS procedure was, under certain assumptions, appropriate to assess the unreliability of measures, but not for detecting invalidity in measurement instruments because only

¹⁸ Because of a strong correlation between the responses to the question on dismissal and the question on promotion.

¹⁹ Congenetic measures are defined as measures of the same underlying construct in possibly different units of measurement or possibly different degrees of precision (Lord and Novick 1968, pp. 47–50). The congenetic measures are indicators of different characteristics that are all simple linear functions of a single construct (Groves 1989, p. 18).

one method was used (Saris and Andrews 1991). Jöreskog (1971) proposed the RMM approach and suggested that measurement instruments could be evaluated with respect to validity by a specific *congeneric*²⁰ test model.

10.3.2.1 How to Disentangle Measurement Validity and Invalidity

Central to the RMM approach is the notion of an unobserved theoretical construct that the researcher attempts to measure by means of observed indicators (questions) characterized by specific response scales. Validity refers thus explicitly to the extent to which an observed measure reflects the underlying theoretical construct that the investigator intends to measure (Cronbach and Meehl 1955, p. 282). Measurement instruments that reflect only weakly the underlying (and intended) constructs because they are systematically affected by other sources of variation are invalid. Reliability is associated with the stability of measures (given that there is no real change). It is perfectly possible to obtain stable measures of measures that are invalid.²¹ For that reason it is recommended not to use mathematical equations that express validity completely in function of reliability. This recommendation is applied in Saris version of the "true score" MTMM model that is shown in the Appendix (Saris 1995) in which it is, contrary to the classic "true score" model (see Lord and Novick 1968), possible to estimate reliability independently of validity when this model is used in a *multitrait-multimethod* context (MTMM).

The MTMM design was at the end of the 1950s developed by Campbell and Fiske (1959) who focused mainly on the assessment of validity. They suggested measuring each of a number of traits with a number of *different* methods in order to build a MTMM correlation matrix. The analysis technique they developed

was largely qualitative and provided a raw indication of the 'convergence' and the 'discriminant' validity of the MTMM measures in view of potential method effects. In reaction to problems that emerged from Campbell and Fiske's strategy of analysis of MTMM correlation matrices, a path-analytic approach to the MTMM matrix introduced the idea of decomposition of the correlations (for covariance) between the measurements (Kallenberg and Kluegel 1975, pp. 2-4) in terms of validity, invalidity, and unreliability parameters. The development of *Structural Equation Modeling* (SEM) software facilitated the application of these models, and Confirmatory Factor Analysis (CFA) became the main method for analyzing MTMM data (Schmitt and Stults 1986). This procedure was most appropriate for evaluating the validity and reliability of measurement instruments; it provides numerical estimates of these measurement quality parameters (Saris and Andrews 1991). Andrews and his colleagues of the *Survey Research Center* provided a convincing account of the advantages of MTMM models for survey research (Andrews and Withey 1976; Andrews and Crandall 1976). The proposed repeated measurements of several traits with different methods in order to obtain estimates of the measurement quality of survey data, and to arrive at 'rules' about how survey questions should best be worded (Andrews 1984).

In the classical MTMM matrix, an observed score y_{ij} is linear decomposed into three unobserved components²²:

$$y_{ij} = v_i F_i + m_j M_j + e_{ij}$$

where y_{ij} is a measure of i th trait by j th method, F_i is the i th trait, M_j the j th method factor, and e_{ij} is random (residual) error. The slope parameter v_i can be considered the validity coefficients, and the slope parameter m_j is the invalidity coefficients. This formulation makes explicit that

²⁰ See footnote n°7.

²¹ Take, for example, an IQ test in Dutch intended to measure the IQ of French-speaking students. The results may be very stable but it measures very probably the students' knowledge of Dutch, and not their IQ.

some of the reliable variance is actually invalid since it is induced by the measurement instrument (or scale) used (Alwin 1997, p. 328). The observed (co)variances in an observed MTMM matrix are thus conceived as the result of two kinds of underlying latent variables: traits and method factors. The residual variance is not explained by the common trait and method factors.

One should notice that in this linear regression approach²³ to measurement the variances in the observed indicators are conceived as 'caused' by the variance in a latent variable, which is conceived as the common source. The strength of a relation between an observed indicator and the latent variable behind indicates "measurement validity" of this indicator. This does not coincide with "theoretical validity" (or construct validity) that relates to the not directly measurable relation between the latent variable and the intended concept.

This MTMM approach was further developed by Saris²⁴ and his research team in an MTMM project that intended to develop strategies for evaluating the quality of survey measures (Saris 1990a, pp. 52-80). An MTMM 'true score' model with measurements of at least three traits each measured by at least three different methods was proposed, and problems with the design were discussed (Saris 1990b, pp. 118-129; De Wit 1994; Scherpenzeel 1995; De Wit and Billiet 1995). The modified MTMM 'true score' was developed in order to be able to estimate the reliability parameters independently from the residual variance (see Saris and Andrews 1991).

²³ Contrary to standard linear regression, the parameter on the left side of the equation is a measured quantity and is expressed in function of un-measured variables on the right side of the equation. For that reason, a set of equations need to be solved in order to obtain values of the regression parameters in covariance structure analysis. The number of unknowns is reduced by the introduction of constraints between the parameters on theoretical ideas, e.g., equality of slopes between indicators and a common method factor.

²⁴ Saris established in 1989 the *International Research Group on Measurement Error in Comparative Survey research* (IRMCS).

10.3.2.2 How to Adjust Measures for Invalidity

There are two ways for adjusting the correlations between traits for measurement invalidity and unreliability, depending on the question whether the quality estimates are directly available in the collected data or not. In the latter, one must derive the estimates from other sources (Andrews 1984, 1990; Scherpenzeel and Saris 1997).

Researchers who repeatedly measured several traits with different response scales, and who analyzed this MTMM data as is the case in ESS, directly obtain correlations between the traits $\rho(F_i, F_j)$ that are adjusted for unreliability and for effects of the methods used. They can apply the most reliable and valid measurement of the traits (see Appendix). In ESS the MTMM design is combined with a (randomized) split ballot design in order to diminish the burden on interviewers and respondents of repeated measurements. Each respondent only received two versions of the experimental questions. This design nevertheless allows the estimation of a full MTMM matrix (Saris et al. 2004). The main problem is the very limited applicability of this direct method because repeated measurements are only available for a small number of measured traits (or variables) in a survey.

Researchers who cannot rely on direct MTMM-based validity and reliability estimates

are advised to publish their correction coefficients, which are based on a meta-analysis of a large number of MTMM studies (Scherpenzeel and Saris 1997). According to these scholars, the average reliability and validity of a measurement should be adjusted depending on the effects of a number of characteristics of the actually used measurement instrument (i.e., subject or domain of the study, response scale, data collection mode, position in the questionnaire, language group, etc.). For development of the core questionnaire and the rotating models, ESS applied the *Survey Quality Predictor* (SQP) program²⁵ (Saris and Gallhofer 2007) to estimate the reliability and validity of the proposed question wordings and response scales. SQP is based on the database of quality estimates that are derived from the growing number of MTMM studies it reports the expected validities and reliabilities of the planned questions.

10.3.3 Measurement of Response Styles

Response style refers to a systematic tendency to respond to a range of survey items on some basis other than what the items were specifically designed to measure (Paulhus 1991). It is a kind of measurement error typical for sets of items characterized by the same format, which is typical for multiple indicators (*congeneric* measures) that measure a latent variable. Four kinds of response styles are distinguished: (1) Agreeing (ARS); the tendency to endorse assertions independently from their content called "acquiescence"; (2) Denying (DRS); the tendency to deny assertions independently from their content (DRS) which is the opposite of yes-saying; (3) Extreme responding (ERS); the tendency to choose extreme response categories of response scales independently from the content of items (ERS); (4) Midpoint responding (MRS); the tendency to choose the middle of a response

scale independently from their content. Response styles might be related to satisficing.

The main challenge in assessing response style is to measure it *independently* from content. Weijters (2006) distinguished two dimensions in the classification of response style measures: the status of the items on which the response measures are based, and the treatment of the content. According to the status, one can distinguish multi-functional measures that measure both content and style by a common set of measures, and specific measures that only measure style. There are four ways of treating the content in response style measures: (a) the content is not deliberately planned or selected before data collection (no specific ex ante control); (b) the content has been eliminated with the aim to measure style; (c) the content is manipulated (e.g. opposite meanings in sets of items) in order to cancel out the effect of content; (d) content is randomized in such a way that systematic influence of content on response is hindered. The classification of *status* by *treatment* does not lead to eight realistic kinds of response style measures since some combinations are per definition impossible (Weijters 2006, pp. 33–43).²⁶

There are attempts to develop content free items and to measure ARS by specific measures, but these were not very effective (Rorer 1965). Random selection of items in view of specific style measures is often used in marketing research. Some scholars (Baumgartner and Steenkamp 2001; Weijters 2006), follow the method proposed by Greenleaf (1992a, b), and make use of a set of items that is maximally heterogeneous in content. The basic idea behind this approach is to reduce the effect of content in the set of items to random noise: if all the items represent different constructs that are unrelated,

²⁶ E.g. the elimination of content in a multi-functional measure, or the randomization of content in multi-functional items that are intended to measure content and style. If one eliminates the content then there is no longer a multi-functional measure, and the selection of items to measure content is per definition not random.

²⁵ See <http://www.sqp.nl/media/sqp-1.0/>.

Fig. 10.1 Measurement model for two interrelated content factors and a style factor

it can be expected that there is no consistency in responses other than that induced by response styles (Weijters 2006, p. 42). Response styles are then measured by the sum of items characterized by specific answers (agreeing, disagreeing, extreme responses, selecting the middle of response scales). The disadvantage of this approach is that a large extra set of unrelated items is needed in order to account for a style effect. The advantage is that it applies to the four response styles.

Multi-functional items combined with experimental control are possible for measuring ARS and DRS when balanced sets are used to measure content. Following a study by Mirowsky and Ross (1991) and a proposal by McClendon (1992), Billiet and McClendon (2000) measured acquiescence for two balanced sets of items. They applied a model with two content factors (ethnic threat and distrust in politics) each measured by a balanced set of

items, and an additional style factor (acquiescence) which loaded equally on all items.

This model yielded a significantly better fit than a model with only two content factors (see Fig. 10.1²⁷ for a formal visualization) or other model specifications.²⁸

As was expected, the two content factors did not correlate with the style factor. This confirms the theoretical postulation about a zero relation between style and content. The study provided a strong support for the existence of acquiescence since the latent variable correlated very strongly

²⁷ The model contains two content factors, Threat and Distrust, with balanced sets of six and four observed indicators, respectively, and with identical slopes for the style factors (denoted 1). For actual parameter values, see Billiet and McClendon 2000, pp. 622–625).

²⁸ Models with two positive and two negative factors, or models with a separate style factor per content.

with the sum of agreements,²⁹ and was replicated in different populations (Billiet et al. 2002).

This approach does not need a separate measurement of style and, contrary to MTMM, no repeated measurements with identical items with different response scales. Content and style are clearly separated in the modeling. The conditions are that at least one more or less congenic set of indicators is used to measure a content variable, and that the items have a specific response scale (Likert scales, or any "agree-disagree" scales). This method is however only applicable to measure ARS (or eventually DRS) and not applicable to the other styles. The method is also applicable to response scales that are considered categorical (or ordinal) in the context of latent class analysis (Moors 2004).³⁰

10.4 Concluding Discussion

This chapter focused on some aspects of two sources of systematic error in surveys: non-response error and measurement error. The structure of this chapter in two parts reflects the separate development of two research traditions, the one on errors of non-observation, the other on observation error. In the first part, only the quality of the obtained response was discussed with focus on non-response reduction and bias estimation. Another error source of non-observation, non-coverage of part of the population in the sampling plan, was not touched. In this part, both the reducers approach (response enhancement) and the measurers (estimation of non-response and bias) were discussed. The response enhancement section was mainly, but not exclusively, covered

during the process of data collection that, in line with a responsive design (Groves and Herring 2006), the completion of the planned sample is permanently fed by information about response expected propensities according to demographic and territorial variables (Schouten et al. 2009).

Adjustment for bias by means of propensity score weightings is an alternative approach. The success of both approaches (R-indicator and adjustment by weighting) depends strongly on the availability of information about variables that are related with the target variables. This information might be at both the respondent level, or at the context level. The likelihood that such information is available depends on the accessibility of databases, which is more likely at group (context) level than at individual level. If relevant context data is available, hierarchical models might be useful. Another future development is the collection of paradata in view of the adjustment of biased sample statistics (Kreuter et al. 2010).

Future research should also focus on sequential mixed-mode designs that are developed to reduce both non-response and survey costs (Dillman et al. 2009). Researchers start with the cheapest mode (e.g. mail or Internet) and then switch to telephone and finally to the most expensive mode (face-to-face) (see Chap. 21 by Shine and Dulisse).

Contrary to the impression one may have at bases of separate treatment of response and measurement error in this chapter, both are actually related since both error components

have an effect on the response distributions. This leads mostly to under-, or overestimation of means and percentages of specific categories of variables, and it may in principle effect correlations or regression parameters. Our experience with ESS learns that measurement error mostly has stronger effects than non-response bias, in particular in case of 'true' scores of subjective variables (platonic true values) as opinions, attitudes, values, and choices. Furthermore, non-response and measurement error are related. Practices to reduce survey non-response may bring into the respondent pool more respondents that provide data filled with measurement error (Olson 2006). This is presumably because reluctant respondents are more likely to satisfy and due to their lower cognitive ability (Kaminska et al. 2011). Moreover, in the actual discussion on the implementation of sequential mixed mode surveys the likelihood of mode effects, a kind of measurement error, is real (Dillman et al. 2009). Researchers who are concerned with response enhancement in surveys should also pay attention to error-sensitive characteristics of additional respondents.

10.5 Appendix

The MTMM true-score model for two correlated variables measured with two different methods (response scales), and with unique components and random error

²⁹ In order to obtain validation, Billiet and McClelland (2000, pp. 623–626) investigated the relation between the style factor and a variable that measures the sum of agreements (named "scoring for acquiescence"), across a balanced set of 14 positively and negatively worded items. This procedure led to a strong correlation of 0.90 ($t = 22.26$) between "scoring for acquiescence" and the style factor in a random sample of Flemish voters in Belgium. This finding was replicated several times in other samples.

³⁰ One can model MRS and ERS with latent class analysis by specifying a latent class of respondents that are very likely to choose the middle of scales or extreme response categories.

y_{11} and y_{22} : observed indicator of i measured with methods 1 and 2;

F_1 and F_2 : state of latent variables 1 and 2 which are correlated;

M_1 and M_2 : latent method factors 1 and 2 that are assumed un-correlated;

T_{11} and T_{22} : true score of variable F_i repeatedly measured with methods 1 and 2;

e_{11} and e_{22} : random measurement error for item i measured with methods 1 and 2;

r_{11} and r_{22} : the reliability coefficients for item i measured with methods 1 and 2; the square of these are estimates of the test-retest reliability in case of repeated identical methods)

v_{11} and v_{22} : the true score validity coefficients for item i measured with methods 1 and 2; v_{11}^2 (or v_{22}^2) is the variance in T_{11} (or the T_{22}) explained by the variable F_i that one intends to measure;

m_{11} and m_{22} : the method effect on the true score of methods 1 and 2; m_{11}^2 (or m_{22}^2) is together with u_{11} (or u_{22}) the part in T_{11} (or T_{22}) not explained by the intended variable, and thus invalidity;

u_{11} and u_{22} : the residual variance of the true score of item i measured with methods 1 and 2;

$\rho(F_1, F_2)$: the correlation between the two latent (intended) variables

$$y_{ij} = r_{ij}T_{ij} + e_{ij} \quad (10.1)$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j + u_{ij} \quad (10.2)$$

This model is a simplified version of the true score MTMM model used. It is in reality not possible to estimate a "2(trait) $\times$ 2(methods)" model. One needs at least three traits and three methods. But even then it is not possible to estimate the parameters, unless certain assumptions are made. Most simple (and acceptable) is to assume that the unique residual variances of the true scores (u_{ij}) are zero (Saris 1990b, p. 119; Scherpenzeel and Saris 1997, pp. 344-347).

In line with Eqs (10.1) and (10.2), and assuming zero unique variance u_{ij} , the true score model is written as:

$$y_{ij} = r_{ij} \times v_{ij}F_i + r_{ij} \times m_{ij}M_j + e_i \quad (10.3)$$

Path analysis suggests that the observed correlation between the measures of two traits, both measured with method M_1 , is (Scherpenzeel and Saris 1997, pp. 372-373):

$$\text{Corr}(y_{11}, y_{21}) = v_{11} \times r_{11} \times \rho(F_1, F_2) \times v_{21} \times r_{21} + r_{11} \times m_{11} \times m_{21} \times r_{21} \quad (10.4)$$

From which one can derive:

$$r(F_1, F_2) = [\text{Corr}(y_{11}, y_{21}) - (r_{11} \times m_{11} \times m_{21} \times r_{21})] / (v_{11} \times r_{11} \times v_{21} \times r_{21}) \quad (10.5)$$

This expression is useful for adjusting observed correlations between variables when one can rely on validity, reliability, and identity estimates (e.g. based on meta-analysis of MTMM data).

References

- AAPOR. (2006). *Standard Definitions. Final Dispositions of case Codes and Outcome rates for Surveys* (pp. 46). American Association for Public Opinion Research: http://www.aapor.org/pdfs/standarddefs_4.pdf
- Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality & Quantity*, 23(3-4), 277-331.
- Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. In P. V. Marsden (Ed.), *Sociological Methodology 1992* (pp. 83-118). San Francisco: Jossey Bass.
- Alwin, D. F. (1997). Feeling thermometers versus 7-point scales. Which are better? *Sociological Methods & Research*, 25(3), 318-340.
- Andrews, F. M. (1990). Some observations on meta-analysis of MTMM studies. In W. E. Saris & A. Van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (pp. 15-51). North Holland: Amsterdam.
- Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 48(2), 409-442.
- Andrews, F. M., & Crandall, R. (1976). The validity of measures of self-reported well-being. *Social Indicators Research*, 3(1), 1-19.

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being. Americans' perceptions of life quality*. New York: Plenum Press.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156.

Belson, W. A. (1981). *Design and understanding of survey questions*. Aldershot: Cower.

Beckers, M., & Billiet, J. (2010). Uphold or revoke? A study of question wording in twelve municipal plebiscites in Flanders. *World Political Science Review*, 6(1), article 5, p. 29; <http://www.bepress.com/wpsr/vol6/iss1/art5/>

Biderman, A. (1980). *Report of a workshop on applying cognitive psychology to recall problems of the national crime survey*. Washington, DC: Bureau of Social Research.

Biemer, P., & Caspar, R. (1994). Continuous quality improvement for survey operations: Some general principles and applications. *Journal of official statistics*, 10(3), 307-326.

Biemer, P. P. (2010). Total survey error, design, implementation and evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (1991). *Measurement errors in survey*. New York: Wiley.

Billiet, J. (2011). Godsdiensstige betrokkenheid en de houding tegenover vreemden: Een verdwijning verband? In K. Abs, K. Dobbelaere & L. Vuyt (Eds.), *Nieuwe tijden, nieuwe mensen. Belgen over arbeid, gezin, ethiek, religie en politiek*. Chapt. X (pp. 215-243). Leuven: Lamoo campus & Konink Bouwduinstelling.

Billiet, J., Waterplas, L., & Loosveldt, G. (1992). Context effects as substantive data in social surveys. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 131-147). New York: Springer.

Billiet, J., & McClendon, J. M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4), 608-628.

Billiet, J., Cambé, B., & Welkenhuyzen-Gybel, J. (2002). Equivalence of Measurement Instruments for Attitude Variables in Comparative Surveys taking Method Effects into Account: The Case of Ethnocentrism. In A. Fertig & A. Mrvar (Eds.), *Developments in Social Science Methodology* (pp. 73-96). Ljubljana: FDV.

Billiet, J., Swyngedouw, M., & Waeghe, H. (2004). Attitude Strength and Response Stability of a Quasi-Balanced Political Alienation Scale in a Panel Study. In W. E. Saris & P. Sudman (Eds.), *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error and Change* (pp. 268-292). Princeton: Princeton University.

Billiet, J., Koch, A., & Philippens, M. (2007a). Understanding and improving response rates. In R. Jowell,

C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 113-137). London: Sage.

Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007b). Estimation of nonresponse bias in the European social survey: Using information from reluctant respondents. *Journal of Official Statistics*, 23(2), 135-162.

Billiet, J., Matsuo, H., Beullens, K., & Vehovar, V. (2009). Non-response bias in cross-national surveys: Designs for detection and adjustment in the ESS. *ASR Research and Methods*, 18, 3-43.

Blom, A., Jäkle, A., & Lynn, P. (2010). The use of contact data in understanding cross-national differences in unit-nonresponse. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, T. L. E. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts* (pp. 335-354). Hoboken: Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bradburn, N. M. (1992). What have we learned? In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 315-324). New York: Springer.

Bortnstedt, G. W. (1983). Measurement. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research: Quantitative studies in social relations* (pp. 70-122). New York: Academic Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Cannell, H. (Ed.). (1944). *Gauging public opinion*. Princeton: Princeton University Press.

Converse, J. M. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206-261). New York: The Free Press.

Converse, J. M., & Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Sage University Series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage Publications.

Couper, M. P., & De Leeuw, E. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 157-178). New Jersey: Wiley.

Couper, M. P. (2001). Measurement error in web surveys. *Methodology & Statistics*.

Cronbach, L. J., & Mehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

Davidov, E., Schmidt, P., & Billiet, J. (Eds.). (2011). *Cross-cultural analysis: Methods and applications* (p. 503). New York: Francis & Taylor.

De Wit, H. (1994). *Cijfers en hun achterliggende realiteit. De MTMM-kwaliteitsparameters op hun kwaliteit onderzocht*. Leuven: K.U.Leuven, Departement Sociologie.

- De Wit, H., & Billiet, J. (1995). The MTMM design: Back to the founding fathers. In W. E. Saris & A. Munnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments* (pp. 39-60). Budapest: Eötvös University Press.
- De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, 15(2), 129-142.
- Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
- Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: Wiley.
- Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR), and the Internet. *Social Science Research*, 38(1), 1-18.
- Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge: Cambridge University Press.
- Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, et al. (Eds.), *Measurement errors in surveys* (pp. 393-418). New York: Wiley.
- Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2), 176-188.
- Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, 56(3), 328-351.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Groves, R. M. (2006). Non response rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70(5), 646-675.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.
- Groves, R., & Herring, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A*, 169(3), 439-457.
- Groves, R., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849-879.
- Hense, D. R. (1969). Separating reliability and stability in test-retest-correlation. *American Sociological Review*, 34(1), 93-101.
- Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradigms from a web survey. *Social Science Computer Review*, 21(3), 360-373.
- Hippler, H.-J., Schwarz, N., & Sudman, S. (Eds.). (1987). *Social information processing and survey methodology*. New York: Springer.
- Hollenan, B. (2000). *The forbidden allow asymmetry: On the cognitive mechanisms underlying wording effects in surveys*. Amsterdam: Ed. Rodopi.
- Jabine, T. B., Straf, M. L., Tanur, J. M., & Rourangeau, R. T. (Eds.). (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.
- Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (Eds.). (2007). *Measuring attitudes cross-nationally: Lessons from the European social survey*. London: Sage.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congenic tests. *Psychometrika*, 36(2), 109-133.
- Kallenberg, A. L., & Khogel, J. R. (1975). Analysis of the multitrait-multimethod matrix: some limitations and an alternative. *Journal of Applied Psychology*, 60, 1-9.
- Kariminska, O., McCutcheon, A., & Billiet, J. (2011). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74(5), 956-984.
- Kalton, G., & Schuman, H. (1982). The effect of the question on survey responses: A Review. *Journal of the Royal Statistical Society*, 145(1), 42-73.
- Kreuter, F., & Kohler, U. (2009). Analysing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustment in the European Social Survey. *Journal of Official Statistics*, 25, 203-226.
- Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzi-Rica, T., Casas-Cordero, C., et al. (2010). Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: Examples from multiple surveys. *Journal of the Royal Statistical Society Series A*, 173(2), 389-407.
- Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research* (pp. 29-44). San Francisco: Jossey-Bass.
- Loosveldt, G., Carton, A., & Billiet, J. (2004). Assessment of survey data quality: A pragmatic approach focused on interviewer tasks. *International Journal of Market Research*, 46(1), 65-82.
- Loosveldt, G., & Beullens, K. (2009). *Fieldwork monitoring: Work package 6, Deliverable 5*. (Version 2): <http://www.risq-project.eu>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., & Trevin, D. (1997) (Eds.). *Survey measurement and process quality*. New York: Wiley.
- Lynn, P. J., Beerten, R., Laitho, J., & Martin, J. (2001). *Recommended standard final outcome categories and standard definitions of response rate for social surveys*. Working Paper 2001, 23 October.
- Lynn, P. (2003). PEDAKSI: Methodology for collecting data about survey non-respondents. *Quality & Quantity*, 37, 239-261.
- Lynn, P., & Kaminska, O. (2011). *The impact of mobile phones on survey measurement error* (p. 45). Working Paper Series, Essex: Institute for Social & Economic Research, 2011-2007.
- Matsuoka, H., Billiet, J., Loosveldt, G., & Malnar, B. (2010a). *Response-based quality assessment of ESS Round 4: Results for 30 countries based on contact*
- files (pp. 80). CeSO: Working paper, CeSO/SM/2010-2.
- Marsuo, H., Billiet, J., Loosveldt, G., Berglund, F., & Kleven, Ø. (2010b). Measurement and adjustment of non-response bias based on non-response surveys: The case of Belgium and Norway in European Social Survey Round 3. *Survey Research Methods*, 4(3), 165-178.
- McClendon, M. J. (1991). Acquiescence and recency response-order effect in interview surveys. *Sociological Methods and Research*, 20(1), 60-103.
- McClendon, M. J. (1992). On the measurement and control of acquiescence in latent variable models. Paper presented at the annual meeting of the American Sociological Association. (pp. 24-29). Pittsburgh: Augustus.
- Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2x2 Index. *Social Psychology Quarterly*, 54(2), 127-145.
- Molenaar, N. J. (1986). *Formuleringseffecten in survey-interviews: een nonexperimenteel onderzoek*. Amsterdam: Vrije Universiteit.
- Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, 20(4), 303-320.
- Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, 70(5), 737-758.
- O'Muircheartaigh, C., Gaskell, G., & Wright, D. B. (1995). Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics*, 11(3), 295-307.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego: Academic Press.
- Payne, S. L. (1951). *The art of asking questions*. Princeton: Princeton University Press.
- Rorer, L. G. (1965). The great response style myth. *Psychological Bulletin*, 63(3), 129-156.
- Rossi, P. H., Wright, J. D., & Anderson, A. B. (1983). *Handbook of survey research: Quantitative Studies in Social Relations*. New York: Academic Press.
- Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly*, 5(1), 91-92.
- Saris, W. E. (1990a). Models or evaluation of measurement instruments. In W. E. Saris, & A. van Meurs. (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (pp. 52-80). Amsterdam: North Holland.
- Saris, W. E. (1990b). The choice of a model for the evaluation of measurement instruments. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (pp. 118-132). Amsterdam: North Holland.
- Saris, W. (1995). Designs and models for quality assessment of survey measures. In W. Saris & A. Munnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments*. Eötvös University Press: Budapest.
- Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-598). New York: Wiley.
- Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating quality of measurement instruments: The split-ballot MTMM Design. *Sociological Methodology*, 34(1), 311-347.
- Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken: Wiley.
- Saris, W. E., & Sniderman, P. M. (Eds.). (2004). *Studies in public opinion: Attitudes, nonattitudes, measurement error and change*. Princeton: Princeton University.
- Scherpenzeel, A. (1995). Misspecification effects. In W. E. Saris & A. Munnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments* (pp. 61-70). Budapest: Eötvös University Press.
- Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, 25(3), 341-383.
- Schmitt, N., & Stults, D. N. (1986). Methodology review. Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10, 1-22.
- Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113.
- Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, 44, 692-712.
- Schuman, H. (1982). Artifacts are in the mind of the beholder. *The American Sociologist*, 17, 21-28.
- Schuman, H. (1984). Response effects with subjective survey questions. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 1, pp. 129-147). New York: Russell Sage Foundation.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context*. New York: Academic Press.
- Schuman, H., & Ludwig, J. (1983). The norm of even handedness in surveys as in life. *American Sociological Review*, 48(1), 112-120.
- Schwarz, N., & Hippler, H. J. (1987). What response scales may tell your respondents. In H. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 163-178). New York: Springer.
- Schwarz, N., Strack, F., Mueller, G., & Chassin, B. (1988). The range of response alternatives may determine the meaning of questions: Further evidence

- on informative functions of response alternatives. *Social Cognition*, 6(2), 107-117.
- Schwarz, N., & Hippler, H. J. (1991). Response alternatives: The impact of their choice and presentation order. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 163-178). New York: Wiley.
- Schwarz, N., & Sudman, S. (Eds.). (1992). *Context effects in social and psychological research*. New York: Springer.
- Schwarz, N., Knipfer, B., Oyserman, D., & Stich, C. (2008). The psychology of asking questions. In E. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology*. (pp. 18-34). New York: Lawrence Erlbaum Associates.
- Smith, T. W. (1984). Nonattitudes: A review and evaluation. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 215-255). New York: Russell Sage Foundation.
- Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology*, 22, 377-399.
- Sniderman, P. M., & Hagendoorn, L. (2009). *When ways of life collide: Multiculturalism and its discontents in The Netherlands*. Princeton: Princeton University Press.
- Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H. Hippler, N. Schwarz & S. Sudman (Eds.), *Social Information Processing and Survey Methodology* (pp. 123-148). New York: Springer.
- Strack, F., Schwarz, N., & Wänke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, 9(1), 111-125.
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.
- Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass Publishers.
- Stoop, L., Billiet, J., Koch, A., & Fitzgerald, R. (2010). Improving survey response: Lessons learned from the European Social Survey. Chichester: Wiley.
- Tourangeau, R. (1992). Context effects on responses to attitude questions: attitudes as memory structures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 35-48). New York: Springer.
- Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103(3), 299-314.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- Vehovar, V. (2007). Non-response bias in the European Social Survey. In G. Loosveldt, M. Swyngedouw & B. Cambré (Eds.), *Measuring meaningful data in social research* (pp. 335-356). Leuven: Acco publisher.
- Weijters, B. (2006). *Response styles in consumer research*. Belgium: Vlerick Leuven Gent Management School.
- Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

Why People Agree to Participate in Surveys

Gerald Albaum and Scott M. Smith

11.1 Introduction

Why do some people participate in survey research as respondents while others fail to participate? There is very little known about the reasons for or against participation and among some respondents why the quality of data varies. A great number of empirical studies have been conducted on response inducement techniques and other methodological artifacts that affect response to surveys, particularly mail surveys. Of concern in these studies have been the effects of devices such as preliminary notification, foot-in-the-door technique, follow-ups, questionnaire format and length, survey source or sponsorship, personalization, cover letters (or e-mails), deadline date, and premiums and rewards. Nature of return envelopes and typed postage are relevant only for mail surveys. Reviews by Peterson and Kerin (1981) and by Yu and Cooper (1983) included all major methods of survey data being used at the time (mail, telephone, personal interview), as did the quantitative review of response effects to interviews by Sudman and

Bradburn (1974). The consensus of all this research is that there is no strong empirical evidence that any technique or aspect of a technique is always best at inducing response, except, perhaps, for pre-notification, follow-ups, and the use of a monetary incentive, (particularly when the incentive that is sent with the request for participation). Typically survey response studies use one or more of the following as a dependent variable or measure of effect: response rate, item omissions, response speed, cost, and in a much more limited manner, response content itself.

Since the late 1990s, online survey data collection has been growing in importance and popularity. Summaries of the research on methodological characteristics of this mode of survey data collection are presented by Miller (2006), Smith et al. (2006), and Roster et al. (2007). Internet-based survey data collection is fast becoming the preferred mode for conducting survey-based research in all areas of business and society.

The lack of a generally accepted framework or theory of survey response has been a hindrance to the development of effective survey techniques. A theory of survey response could explain behavior such that a researcher could predict which techniques would do what under differing conditions. A theory of survey response will guide researchers and help determine how to encourage response and accurate and truthful reporting of data.

This chapter addresses these issues by presenting a framework for explaining survey response that is applicable to all modes of data

G. Albaum (✉)
Robert O. Anderson Schools of Management,
University of New Mexico, Albuquerque, NM, USA
e-mail: albaum@mgmt.unm.edu

S. M. Smith
Brigham Young University, Provo, UT, USA
Qualtrics, PO Box 970340, Orem, UT 84097-0340,
USA