Research into questionnaire design

A summary of the literature

Petra Lietz
Australian Council for Educational Research

Some consider responding to survey questions as a sophisticated cognitive process whereby respondents go through, often iterative, steps to process the information provided to them by questions and response options. Others focus more on the interplay between questions and answers as a complex communication process between researchers and respondents, their assumptions, expectations and perceptions. In this article, cognitive and communication research is reviewed that has tested the impact of different question and answer alternatives on the responses obtained. This leads to evidence-based recommendations for market researchers, who frequently have to make decisions regarding various aspects of questionnaire design such as question length and order, question wording, as well as the optimal number of response options and the desirability or otherwise of a 'don't know' option or a middle alternative.

Introduction

Cognitive research into survey methodology starts from the premise that responding to survey questions involves many, frequently iterative, steps of complex information processing (Cannell *et al.* 1981; Hippler *et al.* 1987; Tourangeau *et al.* 2000; Aday, 2006). The process begins with the comprehension of the question and proceeds to the retrieval of relevant information from memory. Next, it involves a judgement and estimation process that is related to the respondent's motivation and preparedness to be truthful. Ultimately, the respondent's internally generated answer is matched to the response categories provided in the questionnaire.

Others (Hunt *et al.* 1982; Foddy 1993) consider a survey as a complex communication process whereby the product of the interaction between researchers and respondents leads to the sharing and creating of meaning. First, researchers have to agree about what to ask within a framework

Received (in revised form): 2 April 2009

DOI: 10.2501/S147078530920120X

or model encompassing the research questions and hypotheses to be addressed and tested by the information obtained from the study (see also Vikat *et al*. 2007). Second, researchers or interviewers encode their request for information in a carefully standardised physical stimulus, the question, at the beginning of the process. Respondents subsequently decode this stimulus and encode an answer, which is usually expressed in terms of a standardised format, which was previously encoded by the researcher. Finally, the researchers decode this response and proceed to analysing the information and drawing conclusions from the analyses.

Although with somewhat different emphases, both models draw attention to the fact that even minor details in the formulation of questions and answers can have a major effect on the responses obtained and ultimately on the conclusions drawn from the research. Hence, in this article, research into questions and the various possibilities of their encoding and decoding is reviewed first, followed by a discussion of studies investigating the impact of different forms of encoding and decoding responses.

Questions

Brace (2004) has emphasised the importance of question encoding to the success of the communication process, particularly in market research, which has to be able to successfully tune in to the language of respondents that are diverse in terms of gender and age, as well as level of education, occupation and income. Therefore, the research reported below focuses on best practice as regards question length, question wording and question order, in order to avoid negative impact on sample quality due to non-response – which has been shown to increase over time (deLeeuw & deHeer 2002) – or on data accuracy due to respondents' misinterpretation of or deliberate lying in answer to questions. It should be kept in mind that good practice in terms of these issues is of particular importance in international research as it assists in reducing the impact of difference in culture and language on survey results (Brislin 1986; Smith 2003).

Question length

Results of a multi-level analysis undertaken by Holbrook *et al.* (2006) confirm the general advice to keep questions or statements as short as possible (Foddy 1993; Dillmann 2000; Fink 2003) in order to increase respondents' comprehension. For the English language, Brislin (1986) specifies a maximum number of 16 words, while Oppenheim (1992)

recommends 20 words per sentence whereby questions can consist of more than one sentence.

In addition, Blair *et al*. (1977) and Andrews (1984) report increased data quality if questions or groups of questions concerning the same topic are preceded by a medium-length introduction (30 words, Blair *et al*. 1977; 16 to 64 words, Andrews 1984). According to evidence reported by Oksenberg and Cannell (1977, p. 342) and Jabine (1987), somewhat longer questions lead to more accurate reporting as they may convey the idea that the task is important and deserves serious effort.

Grammar

Various authors (Brislin 1986; Dillman 2000; Dörnyei 2003) argue to keep the grammatical complexities to a minimum. Thus, questions should employ the active rather than the passive voice, repeat nouns instead of using pronouns and avoid possessive forms. In this way, cognitive demands on respondents are minimised and mental capacity is freed up in order to think about a response.

Specificity and simplicity

Another means of reducing the cognitive load on respondents stems from using specific rather than general terms (Brislin 1986; Dillmann 2000; Martin 2002; White *et al*. 2005), breaking down more complex questions into simpler ones (Jobe & Mingay 1989), providing behavioural illustrations of certain concepts (e.g. that a 'chronic' health condition means seeing a doctor two or three times for the same problem in the last 12 months; Fowler 2004) and avoiding words that indicate vagueness, such as 'probably', 'maybe' or 'perhaps' (Brislin 1986; Dillmann 2000).

Results of multilevel analyses reported by Holbrook *et al*. (2006) have shown that the level of abstraction and the request for qualified judgements increase comprehension difficulties whereby these difficulties are not dependent on age or education as is frequently assumed. Similarly, Belson (1981) and Foddy (1993) advise against the use of hypothetical questions concerning respondents' future behaviours. Instead, it is recommended to use vignettes or alternative scenarios when seeking reactions to issues that are outside the realm of the past or present.

Many studies (Oksenberg & Cannell 1977; Rockwood *et al*. 1997; Tourangeau *et al*. 2000; Schaeffer & Presser 2006), as well as the meta-analyses of Sudman and Bradburn (1974) and Bhandari and Wagner

(2006), show that the invalidity of responses due to cognitive overload increases where recall of events is involved that have occurred not in the immediate past (i.e. more than a week/month ago) whereby the invalidity of responses depends on the importance of the event (e.g. visit to the GP vs hospitalisation; minor vs major house repairs).

Social desirability (SD)

The merit of simplicity in question wording is emphasised by Foddy (1993), who labels the undesired off-putting effect of poorly worded questions on respondents 'question threat'. He adds that the use of difficult vocabulary either in questions or instructions leads to respondents feeling stupid or uneducated, and increases the probability of obtaining 'don't know' or socially desirable responses. Socially desirable responses can lead to answers that inaccurately reflect respondents' actual behaviours in a number of ways. First, respondents might choose to select a certain position that is thought to be one that is favoured by society (e.g. not to smoke or drink, to do exercise). As a consequence, particularly in medical research, people tend to under-report unhealthy lifestyle practices and over-report healthy ones (Brace 2004). Second, because of the social prestige that is attached to the act of uttering an opinion and the corresponding negative evaluation associated with the lack thereof (Leverkus-Brüning 1966), respondents think that they should be informed about certain issues (e.g. the EU constitution, climate change) and give responses conveying this impression instead of admitting ignorance. Third, Foddy (1993) states fear of being identified or revealing details about the private sphere or facts that are considered embarrassing, such as medical diagnoses of mental or genital diseases (Oksenberg & Cannell 1977), as reasons for respondents giving socially desirable responses. It is mainly the first two aspects that are subsumed in Holtgraves' (2004, p. 161) definition of social desirability, which 'refers to a tendency to respond in self-report items in a manner that makes the respondent look good rather than to respond in an accurate and truthful manner'.

In order to reduce respondents' propensity to give socially desirable answers, especially on sensitive issues such as adultery, crime or drug use, Brace (2004) suggests indirect questioning, such as 'What do you believe other people think about …', whereby the assumption is that respondents will more easily admit to views or behaviours that they think are not shared by the larger society by projecting their own views onto others. Or, if the issue involves knowledge that the respondent might not have, a phrasing

such as 'Have you had time yet to familiarise yourself with the new (EU) Constitution?' might facilitate the respondent's acknowledgement of his/her ignorance in this matter.

Another means of reducing respondents' propensity to give socially desirable answers is the use of the introductory phrase 'Do you happen to know …', as Brace (2004) argues that this phrase allows respondents to think a bit longer in order to retrieve any knowledge they might have regarding the topic. Another beneficial aspect of this phrase is put forward by Bradburn *et al.* (2004), who suggest this question wording in order to signal to participants with less firm attitudes or information bases that it is acceptable to volunteer a 'don't know' response.

Other suggestions to reduce social desirability frequently include that questions (a) are worded as neutrally as possible, (b) propose values on a certain topic not only in one but different directions, and (c) suggest the normalcy of socially deviant behaviour (Oppenheim 1992; Bortz & Döring 2003; Scholl 2003; Brace 2004). Diekmann (2003), however, has reported limited effects of such measures.

In addition, a number of instruments (e.g. Edwards Social Desirability Scale, Edwards 1957; Marlowe-Crowne Social Desirability Scale (MCDS), Crowne & Marlowe 1960; Balanced Inventory of Desirable Responding (BIDR), Paulhus 1984) have been developed to measure SD in order to control for this tendency in subsequent analyses (Seitz 1977; Diekmann 2003). However, as much research has reported questionable validity and reliability (Paulhus & Van Selst 1990; Paulhus 1991; Moorman & Podsakoff 1992; Leite & Beretvas 2005) for these instruments, it seems that, although not much research has been done to test empirically the differences in SD that are likely to exist between countries (Stocké & Hunkler 2007), the proposed question wordings that are aimed at reducing the tendency to give socially desirable answers are preferable to the use of measures of social desirability that have questionable psychometric properties and would increase considerably the length of a questionnaire.

Double-barrelled questions

A number of authors recommend avoiding the ambiguity of so-called 'double-barrelled' questions or statements that contain two different verbs or two different concepts. More specifically, Brislin (1986) mentions the use of two verbs in one question as being detrimental to the obtaining of accurate responses, while Fowler (1992), van der Zouwen (2000), Fink (2003) and Brace (2004) extend it to the use of two concepts in

one question. For example, the question 'Do you have time to read the newspaper every day?' contains two aspects, namely 'having the time' and 'reading the paper every day', which is why the question 'Do you read the newspaper every day?' followed by a question about reasons if this is (not) the case will be clearer. This question also illustrates that questionnaire designers have to be clear what it is that they want to obtain information on. At the start, the questionnaire designer might not have realised that the question contained two aspects, namely the behaviour and the reason for the behaviour. On a somewhat different aspect of double-barrelledness, a question such as 'Should older people who smoke pay some of the costs related to a potential lung-cancer treatment themselves?' leaves open who the reference group is. Older people who do not smoke? Younger people who smoke? Younger people who do not smoke?

Negatively worded questions

The general advice is against the inclusion of negatively worded questions or statements (Belson 1981; Foddy 1993) as they have been found to take longer to process (Weems *et al.* 2002) and have a greater likelihood of respondents making mistakes (Eifermann 1961; Dudycha & Carpenter 1973), thus introducing an artificial methods effect into the response behaviour (DiStefano & Motl 2006). Foddy (1993) argues that this is particularly the case when the word 'no/not' is used together with words that have a negative meaning. Thus, he suggests that the question 'What is your view about the statement that conservationists should not be so uncooperative with the government?' should be rephrased into 'What is your view about the statement that conservationists should cooperate with the government?' so that respondents do not have to go through a tiring process in order to deduce the meaning of the question. In addition, he emphasises how quickly a question can turn into a double negative when taken together with the answer options, as is the case when respondents are asked to agree or disagree with the statement 'Teachers should not be required to supervise students in the halls.' O'Muircheartaigh *et al.* (2000, p. 22) confirmed the undesirability of negatively worded items as their analyses showed these to be less reliable than positively worded items. This evidence supports the notion that the introduction of negatively worded items into an item battery in order to balance it 'introduces greater random error' although there is some evidence that this may not be the case if sophisticated item response techniques are used in the development of the scale (Bergstrom & Lunz 1998). An interesting aside in this context

is the finding that more people are willing to respond 'no' to 'allowing' something (e.g. x-rated movies, cigarette advertisements) than to respond 'yes' to 'forbidding' it (Schumann & Presser 1977, 1978; Hippler & Schwarz 1986).

Adverbs of frequency

Another recommendation for clear question wording concerns the use of adverbs that indicate frequency. In an early study, Simpson (1944) asked people for 20 frequency adverbs to indicate the percentage of time this word meant that something occurred. He found the largest agreement for the terms 'never' (0–2% of the time), 'almost never' (3–5% of the time), 'about as often as not' (48–50% of the time) 'always' (98–100% of the time), and the largest difference in interpretation for the terms 'frequently' (40–80% of the time) and 'rather often' (45–80% of the time). Moreover, he found no frequency terms that were interpreted by people to indicate occurrences of between 20 and 50% of the time. Similarly, Liechtenstein and Newman (1967) reported the smallest range in interpretation for the 'middle-of-the-frequency-road' term 'tossup' (45–52%) and the largest range in interpretation for the terms 'predictable', 'probable' and 'possible' (all from 1–99%).

Since then, a general consensus has emerged that 'frequently', 'usually' and 'regularly' have quite different meanings for different respondents, and depending on the question content (Bradburn & Miles 1979; Krumpal *et al*. 2008) as well as on the numeric values assigned if these terms are used as labels of a response scale (Schwarz *et al*. 1998). To highlight this, Foddy (1993, p. 43) reported 445 interpretations of the word 'usually', as the meaning assigned to the word varied depending on, for example, the type of activity or who was asked about the activity.

One solution to this problem is to offer participants more specific quantifiers in the response options. Therefore, 'never or almost never', 'once or twice a month', 'once or twice a week' and 'always or almost always' are used as response options to many of the questions asked in background questionnaires addressed at teachers and principals as part of internationally comparative studies in education (e.g. Mullis *et al*. 2007). In addition, as suggested by Bradburn and Sudman (1979), questions aimed at obtaining information regarding frequency of behaviour should include numeric reference points for a specified time period. Thus, a question about watching television should be worded 'How many hours do you watch TV on a weekday (excluding weekends)?' with response

options such as '<0.5 hours', '0.5 hours to <1 hour', '1 hour to <1.5 hours', '1.5 hours to <2 hours', '2 hours to <2.5 hours', '>2.5 hours'. Of course, this requires accurate knowledge about the question topic to enable the appropriate specification of the time period in the question (Dillman 2000; Fink 2003) and the response categories offered as answers (Schwarz *et al*. 1985; Gaskell *et al.* 1994; Bowling 2005).

Question order

Question order effects arise when answering behaviour changes depending on the position of a question during the interview (Schumann & Presser 1996; Baker 2003). They are problematic in that they not only threaten the validity of the results but also the generalisability of results to the population. Types of question order effects include effects of part–whole combinations, part–part combinations and salience.

Question order effects of part–whole combinations occur where one question is more general with respect to a certain concept while the other is more specific. Examples are questions about respondents' state of happiness in general and their happiness in marriage or respondents' views on abortion in general and on abortion for specific reasons. Systematic research into this issue has been inconclusive as regards the answering behaviour in response to specific questions. For the general question, however, results tend to show that the general question is more appropriately placed before the specific question. This is argued to be due to the fact that the specific question takes a certain aspect out of the concept (e.g. marital happiness from general happiness or severe disability for the concept of abortion), which, then, is removed in the respondent's mind if the general question is asked after the specific question (Schwarz & Sudman 1992; Schumann & Presser 1996; Bates *et al*. 2006).

Question order effects of part–part combinations arise where questions are asked at the same level of specificity and respondents adapt their answers as a result of normative consistency. Thus, in two questions on (a) whether or not US reporters should be allowed into what was then the Soviet Union, and (b) whether or not reporters from the Soviet Union should be allowed to enter the US, Schumann and Presser (1996) found agreement with the second question to be significantly greater if (a) preceded (b) than if (b) was asked before (a). The authors reported similar results for questions regarding allowing US citizens to join the British, French or German armies and vice versa, in that agreement to allow foreigners into the US Army was far higher if this question was asked

second. Counter-evidence, however, emerged for experiments regarding the extent to which people thought lawyers or doctors served the public good as well as a question where respondents were asked for their self-placement into a social class before and after questions regarding their education and occupation. In neither case did a question order effect emerge. Thus, it appears that it depends on the topic as to whether or not question order effects arise for part–part combinations.

Question order effects as a result of salience are said to occur when response behaviour changes as a result of a topic having been raised as part of the questioning process, hence conveying the importance of that topic to respondents (Schumann & Presser 1996). Gaskell *et al*. (1994) found that between 9% and 13% more respondents reported annoyance with adverts and feeling unsafe if previous questions in the survey had touched on these topics.

Demographic questions about respondents, such as age, education, income and marital status, should come at the end of the questionnaire rather than at the beginning in order to avoid negative feelings about the provision of personal information impacting on the answering behaviour or participation (Converse & Presser 1986; Oppenheim 1992).

In summary, the evidence on research into question design suggests that questions should be constructed to be as clear, simple, specific and relevant for the study's research aims as possible. This serves two purposes. First it facilitates comprehension of the questions by the respondents. Second, it assists in keeping questionnaires to a manageable length as, for every question, it should be specified how the information gained will be used to test a hypothesis or contribute to answering the research question. In addition, questions should focus on current attitudes and very recent behaviour in order to increase the accuracy of the reported information (Bradburn & Sudman 2003). Then, more general questions should precede more specific questions as the latter have been shown to influence responses to the former, but not vice versa. Finally, demographics questions should be asked at the end of the questionnaire in order not to affect negatively the preparedness of respondents to answer questions due to the feeling of losing anonymity, which could happen if they were asked those questions at the beginning of the questionnaire.

Responses

The second main area for consideration in the survey communication framework revolves around the responses that are given to answer

questions. Here, the relevant issues pertain to the standardised format or response stimuli in the form of response categories or scales generated on the part of the researcher, as well as the process of encoding on the part of the respondent.

Don't know option

Probably the first central issue that needs to be addressed on the part of the researcher is whether all respondents should answer all questions or whether those respondents with little or no knowledge should be filtered out and not be asked certain questions (Oppenheim 1992; Fife-Schaw 2006). A related issue is – in the context of a standardised interview that is conducted in person, either face to face or by telephone – whether response scales should offer a 'don't know' (DK) option, either explicitly or record it only when it is volunteered. To investigate this issue, Schumann and Presser (1996) conducted 19 experiments that compared responses to questions on US foreign affairs, courts, governments and leadership with and without an explicitly offered DK option. They found that the percentage of respondents choosing DK increased by between 22% and 25% when it was explicitly offered, which was in line with findings reported by Trometer (1996). This difference in percentages held regardless of the familiarity of respondents with the question topic as, for example, a question regarding the Portuguese government with which respondents were less familiar increased from 63.2% to 87.9%, whereas the DK proportion in response to a question regarding the US government increased by about the same amount, from 15.2% to 37.6%. Looking at it in a different way, about one-fifth of respondents shifted from the DK option to a substantive response option (i.e. 'agree' or 'disagree') if the DK option was not explicitly offered.

To examine whether or not the explicit offering of a DK option altered the distributions for the substantive response categories, Schumann and Presser (1996) compared the proportions of respondents choosing the agree and disagree options after omitting the respondents who chose the DK option in the two response types. Results indicated a large significant difference regarding respondents' choice of substantive response options for only one of the 19 experiments.

Opinion floating

Schumann and Presser (1996, p. 118) label people who give a substantive response when the DK is not offered, but who choose this option

when it is offered, floaters as these people seem to vary their responses depending on the response options on offer. To investigate the extent to which they may systematically differ from other respondents, the authors conducted further experiments. Their results showed that while, in general, less-educated respondents tended to give more DK responses than more educated respondents, it was the latter group for which a higher percentage of DK was recorded when the question topic had virtually not been covered in the media. The authors argued that this showed that, for topics that were generally less widely known, more educated respondents were willing to admit ignorance, whereas less educated respondents used information given by the question to develop a substantive response. The authors (Schumann & Presser 1996, p. 160) concluded, 'whether filtered or standard questions should be used in a questionnaire would seem to depend on whether an investigator is interested mainly in an "informed opinion" on an issue or mainly in underlying disposition'.

Opinion filtering

A more explicit way of filtering out respondents is to ask questions such as 'Do you have an opinion on this or not?' or 'Have you been interested enough to favour one side over the other?' While such questions are advocated by some as a means of excluding anyone who is 'ignorant' on a particular issue, two things have to be kept in mind. First, respondents' self-identification as being 'ignorant' might vary systematically as a consequence of question topic as well as respondents' characteristics such as gender and age. Second, a serious consequence of filtering out respondents is the impact on the representativeness of the sample, in particular where stronger filter questions are used (e.g. 'Have you already thought sufficiently about XYZ so that you could form an opinion?' instead of 'Do you have an opinion on XYZ?') that lead to the overestimation of people without an opinion (Hippler *et al*. 1987). A commonly used rule of thumb in survey research (Martin *et al.* 2007) is to consider a sample as being not representative of the intended target population if information is obtained from less than 80% of originally selected participants.

Bishop *et al*. (1979) tested the hypothesis that filtering out respondents through specific questions did not make a difference to the magnitude of the correlations between attitude items. To this end, they examined responses to five studies of US adults with comparable sample compositions in terms of age, gender, race and education. Correlation analyses between respondents' attitudes towards government responsibilities, legalisation

of marijuana and their self-reported location on the liberal–conservative continuum showed higher correlations when filtering questions were applied. The authors argued that this evidence supported their 'non-attitude hypothesis', according to which higher correlations should emerge between political attitude items with a prior filter than for items without a prior filter, since the former would exclude respondents without firm attitudes.

Evidence that runs counter to the hypothesis that less well-informed people have no attitudes on certain issues stems from such people being consistent in their response behaviour over time. Moreover, for the group of people with non-attitudes it could be anticipated that half of them would favour an issue and the other half would oppose an issue. However, in an experiment involving questions that asked about issues to which the general public was known to have had little, if any, exposure Schumann and Presser (1996) found that this was not the case. This substantiated the earlier assumption by Allport (1935, as cited in Schumann & Presser 1996) that people used their general attitudes to guide them in the evaluation of questions with unfamiliar content. The experiments also provided supportive evidence for this assumption in that substantive responses to less well-known issues were related in a systematic way to other items that asked about similar issues but whose content was more widely known. This combined evidence led the authors to conclude that 'the evidence … narrows, if indeed it does not eliminate, the conceptual distinction between attitudes and non-attitudes' (Schumann & Presser 1996).

Number of response scale options

A number of authors (Dillman 2000; Mayer 2002; Fink 2003; Brace 2004) report that between 5-point and 7-point scale response options are the most commonly used, with Dawes (2008) finding that 5- and 7-point scales can easily be rescaled in order to facilitate comparisons. The 7-point scale has been shown to be more reliable (Cronbach 1951) as it allows for greater differentiation of responses than the 5-point scale (Finn 1972; Masters 1974; Alwin 1992) while not artificially increasing differentiation (Cox 1980; Schwarz & Hippler 1991; Porst 2000), as might be the case where more scale points are offered.

Other authors also report evidence that supports the use of longer response scales. Rodgers *et al.* (1992), who investigated the effect of scale length from two to ten response options, found that the expected value of the validity coefficient increased by about 0.04 for each additional

response option. Similarly, Coelho and Esteves (2007) reported higher discriminant validity for a 10-point than a 5-point scale when applied to customer satisfaction measurement. However, Matell and Jacoby (1971) found no linear increase when comparing concurrent validity coefficients for scales with 2 to 19 response options. Alwin (1997) conducted a confirmatory factor analysis of concepts being measured on 7-point scales (labelled 'satisfied' to 'dissatisfied' and 'delighted' to 'terrible') compared to concepts being measured by a number of 11-point 'feeling thermometers'. Results indicated that 11-point scales had consistently higher reliability and validity coefficients and lower invalidity coefficients. This higher quality of 11-point scales, preferably with labelled fixed reference points, confirms the results of a meta-analysis conducted by Saris *et al.* (2004), and has also been found to hold for cross-national studies (Fitzgerald & Widdop 2004).

Instead of relating the optimal length of response scales to the distribution of responses, Foddy (1993) relates it to the content of the question. Thus, Foddy argues that shorter scales, such as 5-point scales, are preferable in situations where respondents are asked for absolute judgements. In contrast, he considers longer scales such as 7- to 9-point scales to be more appropriate in situations where more abstract judgments are sought from respondents.

Odd or even number of response scale options

In addition to the question regarding the optimal number of response scale options, a decision has to be made as to whether to offer respondents an even or an odd number of response scale options. This implies a decision on whether or not to offer a – usually neutral – 'middle' option that allows respondents not to commit themselves to a direction in their opinion or attitude.

Much research (Kalton *et al*. 1980; Garland 1991; Schumann & Presser 1996; O'Muircheartaigh *et al*. 2000) has shown that a middle alternative attracts between six and 23% of respondents when it is offered, although, contrary to popular belief, the tendency to choose a middle option is not generally dependent on age, education or gender (Kalton *et al.* 1980). O'Muircheartaigh *et al*. (2000) proceeded to examine in detail the shift in response distribution that occurred as a result of the inclusion or omission of the middle alternative. They found that the omission of the middle alternative increased responses to the DK option only slightly, by 1% to 2%. In addition, results showed a slightly higher increase for the weak agree/disagree responses (8.5%) than for the more extreme agree/disagree

responses (4.1%) if the middle option was omitted. This latter result was also in line with the results of an experiment by Schumann and Presser (1996), who found that the introduction of moderate alternatives in a question about liberal-conservatism (i.e. 'somewhat liberal' and 'somewhat conservative') attracted more respondents from the middle alternative than from the extreme response alternatives.

O'Muircheartaigh *et al*. (2000) also examined the 'satisficing' hypothesis initially put forward by Krosnick (1991). Krosnick (1991) hypothesises that because many survey participants are likely to have low motivation and may find the task of responding difficult and exhausting they select the response alternative that involves the least amount of thinking and justifying. One of the implications of the satisficing hypothesis is the expectation that an omission of the middle alternative results in people 'reporting meaningful attitudes that they would otherwise not have bothered to describe' (O'Muircheartaigh *et al*. 2000, p. 20). Results of O'Muircheartaigh *et al*.'s (2000) analysis, however, did not support this hypothesis. Instead, response scales without the middle point had lower validity and higher random error variance, indicating that people randomly chose other available response options when the middle option was not available. This desirability of having a neutral middle point to increase the reliability and validity of response scales has also been confirmed by a meta-analysis of 87 experiments of question design reported by Saris and Gallhofer (2007).

O'Muircheartaigh *et al*.'s (2000) analyses also revealed some insights into a phenomenon called 'acquiescence' (Lenski & Leggett 1960; Fife-Schaw 2006), which refers to the tendency of respondents to agree with any statement regardless of its content.

Their analyses confirmed other evidence that such an effect exists (Smith 2004), and highlighted that a two-factor model consisting of (a) the actual attitude towards science and technology and (b) acquiescence was the model that fitted the data best.

Labelling of response scale options

Decisions regarding the labelling of response scale options include whether to use numbered scales that are unipolar (e.g. 'On a scale from 0 to 10 …') or bipolar (e.g. 'Consider a scale from –5 to +5 …') or verbal scales (e.g. 'agree, slightly agree, neither agree nor disagree, slightly disagree, disagree' or 'Would you say that you're very happy, pretty happy or not too happy these days?') and whether to label all responses options or only some of the response scale options.

Evidence from a number of studies (Schwarz *et al*. 1991; Fowler 1995; O'Muircheartaigh *et al*. 1995) has shown a greater likelihood for respondents to choose positive ratings on the bipolar numeric scale than ratings of greater than five on the unipolar numeric response scale. This finding held for topics as different as the entertainment value of movies and TV to general life satisfaction.

In addition, O'Muircheartaigh *et al*. (1995) investigated the effect of differential response scale labelling not only in terms of numbers but also verbal anchors. They reported that the explicit mentioning of the verbal anchors made a difference to responses only on the '0' to '10' scale in that the '0' response option was chosen whereas it was not selected when the verbal anchors were not explicitly mentioned.

In a second experiment, O'Muircheartaigh *et al*. (1995) compared four combinations of unipolar and bipolar numerical and verbal scales. First, they found that the midpoint of both numerical scales (i.e. –5 to +5 and 0–10) was chosen far more frequently (by about 30% of respondents) for the bipolar verbal anchors (i.e. the advertising authority should be given 'much less power' and 'given much more power') than the unipolar verbal anchors (i.e. 'not given any more power' and 'given much more power', chosen by about 20% of respondents). Second, the lowest scale points ('0' and '–5' respectively) were chosen far more frequently if the verbal anchors were unipolar (16% and 15% respectively) than when they were bipolar (7% and 6% respectively). These results are in line with those reported by Schwarz *et al*. (1991), who conclude that respondents use the numeric values to disambiguate the meaning of scale levels, resulting in different interpretations – and ultimately scores – depending on the specific combination of numeric and verbal labels used in a particular study.

A number of studies have investigated the verbal labelling of response scales tapping into the 'good–bad' continuum (Mosier 1941; Myers & Warner 1968; Vidali 1975; Wildt & Mazis 1978). Results indicated that the words 'disgusting', 'unsatisfactory', 'neutral', 'desirable' and 'excellent' produced normal distributions that overlapped little, whereas words such as 'acceptable', 'important' and 'indifferent' polarised respondents. In addition, participants with very different backgrounds rated 'fantastic' and 'excellent' (Mittelstaedt 1971; Myers & Warner 1968) to be the most positive adjectives, and 'horrible' and 'terrible' to be the most negative adjectives. Finally, for the term 'delightful', respondents varied the least, whereas for the term 'unacceptable' respondents varied the most.

Other research has investigated the effects of so-called 'multiplying adverbs' or 'intensifiers' (e.g. 'slightly', 'rather', 'extremely') on

response distributions. Thus, Cliff (1959) asked respondents to rate the favourableness or otherwise of adjectives (e.g. 'respectable', 'mediocre') with and without such adverbs. He found that 'slightly' and 'somewhat' had the smallest intensifying effect, 'very' and 'extremely' had the largest intensifying effect, while 'pretty' and 'quite' were closest to the meaning of an adjective without an intensifier. Similarly, Worcester and Burns (1975) found that adding 'slightly' to the two moderate points of a 5-point agree–disagree scale decreased the overlap of answers. O'Muircheartaigh *et al.* (1993) examined the effect of adding (a) 'really' to a question on the frequency of feeling annoyed by an advert on television, (b) 'very' to a question regarding the frequency of feeling unsafe around the neighbourhood in which they live, and (c) 'extreme' to a question on the frequency of experiencing physical pain. While the effect on the distribution of responses for (a) and (b) was negligible, a significant shift in distribution occurred when 'extreme' was added to the question regarding physical pain. Indeed, only 38% of respondents were aware that an intensifier had been used in the question about television advertisements, whereas 75% of respondents were aware of the use of the word 'extreme' in the question regarding physical pain. This could, however, be a result of respondents assigning a much higher intensity to 'extremely' than 'very', as was demonstrated by Bartram and Yelding (1973).

Order and direction of response options

Foddy (1993) has outlined a number of response options effects, including the primacy and recency effect, as well as the effects of shifting frames of reference. The primacy effect refers to the assumption that respondents will select earlier alternatives more frequently than later alternatives, especially when alternatives are presented on 'show cards'. The recency effect is said to apply when respondents select the later alternatives, and is thought to apply mainly when respondents only hear the alternatives. The phenomenon of shifting frames of reference refers to the possibility that the selection of a certain alternative depends on whether the 'more favourable' alternatives are presented earlier or later. Schumann and Presser (1996) examined these effects in detail and found some evidence of a recency effect, but only for unusual topics and long-winded questions, as well as of a primacy effect for very long lists that present 16 alternatives.

A final issue in the construction of response scales is whether or not the direction of response options (i.e. whether 'strongly agree' should be put

on the left-hand side and 'strongly disagree' on the right-hand side, or vice versa) affects respondents' behaviour. Here, Rammstedt and Krebs (2007) showed that the direction did not affect mean scores and standard deviations significantly as long as the 'strongly agree' option corresponded to the highest numerical value (='8') and 'strongly disagree' to the lowest numerical value (='1'). In addition, for only two out of five personality scales, a significantly lower mean was observed when the numerical value was highest for the 'strongly disagree' option and lowest for the 'strongly agree' option. This seems to confirm Fink's (2003) assertion that the direction of the response options is negligible in most situations and that the lowest numerical value should be attached to the disagree option in order not to make the tendency of respondents to disagree lower than it already is. Finally, Bradburn *et al*. (2004) recommend to put those options first (i.e. on the left) that convey less socially desirable responses, to prevent respondents from making a choice without having to read all available options.

Conclusion

While much obviously depends on the aims of the research, the target population and the special context in which a questionnaire is developed, a number of general recommendations emerge from the above review of research into questionnaire design.

- Questions should be constructed to be as clear, simple, specific and relevant for the study's research aims as possible.
- Questions should focus on current attitudes and very recent behaviour.
- More general questions should precede more specific questions.
- Vague quantifiers such as 'frequently', 'usually' and 'regularly' should be avoided. Instead, carefully pre-tested response options should specify the number of times per appropriate period (e.g. day, week, month, year) of an event or behaviour.
- A desirable Likert-type response scale length ranges from five to eight response options.
- The inclusion of a middle option increases the validity and reliability of a response scale slightly.
- The numerical scale should be unipolar with matching verbal labels as anchors at both ends of the scale.
- 'Extremely' and 'not at all' can serve as most effective verbal intensifiers.
- All numeric labels should be shown to respondents.

- Numeric and verbal anchors (=endpoints) should be mentioned explicitly.
- 'Disagree' options should have lower numeric values attached to them than 'agree' options.
- A 'don't know' option should be recorded if volunteered, whereby interview instructions should be such that interviewers are not to encourage respondents to choose a substantive response option if they hesitate.
- Demographic questions should be put at the end of the questionnaire.

Of course, adherence to these recommendations for questionnaire design will only serve to go some way in the development of a questionnaire that is of high quality. The next step in the questionnaire design process will be the cognitive (e.g. Jobe & Mingay 1989; Drennan 2003; Willis 2005) and quantitative piloting (e.g. Presser & Blair 1994; DeVellis 2003; Litwin 2003) of the questionnaire in order to allow for an evaluation in terms of its acceptance and understanding by members of the intended target population, and an analysis of the psychometric properties (e.g. Andrich 1978; Wright & Masters 1982; Nunnally & Bernstein 1994; von der Linden & Hambleton 1997) of its constituent questions and scales.

Finally, the above literature review of empirical studies that have tested particular aspects of questionnaire design has two main implications for research. First, the theories developed in communications and cognitive research provide helpful and somewhat complementary frameworks for conceptualising research into questionnaire design. Moreover, these theories assist in identifying and making explicit the many interrelated aspects involved in the generation and processing of questions and answers that warrant investigation. Second, the review shows that some of these aspects – for example, the effects of differences in the number of response scale points and their labelling – have been quite extensively investigated by empirical observation. Other aspects – such as the arrangement of response scale labels (e.g. 'strongly agree' to 'strongly disagree') from left to right, or vice versa, and the asymmetry of response categories – would benefit from more empirical observation.

References

Aday, L. (2006) *Designing and Conducting Health Surveys: A Comprehensive Guide* (3rd edn). San Francisco, CA: Jossey-Bass.

Allport, G. (1935) Attitudes. In C.M. Murchison (ed.) *Handbook of Social Psychology*. Worchester, MA: Clark University Press, pp. 798–844.

Alwin, D.F. (1992) Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology*, **22**, pp. 83–118.

Alwin, D.F. (1997) Feeling thermometers vs 7-point scales: which are better? *Sociological Methods and Research*, **25**, pp. 318–340.

Andrews, F. (1984) Construct validity and error components of survey measures: a structural modeling approach. *Public Opinion Quarterly*, **48**, 2, pp. 409–442.

Andrich, D. (1978) Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, **38**, pp. 665–680.

Baker M.J. (2003) Data collection – questionnaire design. *Marketing Review*, **3**, 3, pp. 343–370.

Bartram, P. & Yelding, D. (1973) The development of an empirical method of selecting phrases used in verbal rating scales: a report on a recent experiment. *Journal of the Market Research Society*, **15**, pp. 151–156.

Bates, N., Martin, E., DeMaio, T.J. & de la Puente, M. (2006). *Questionnaire Effects on Measurements of Race and Spanish Origin. Research Report Series.* Washington, DC: Statistical Research Division, US Census Bureau.

Belson, W.A. (1981) *The Design and Understanding of Survey Questions*. Aldershot: Gower.

Bergstrom, B.A. & Lunz, M.E. (1998) Rating Scale Analysis: Gauging the Impact of Positively and Negatively Worded Items. Paper presented at the Annual Meeting of the American Educational Research Association, 13–17 April.

Bhandari, A. & Wagner, T. (2006) Self-reported utilization of health care services: measurement and accuracy. *Medical Care Research and Review*, **63**, 2, pp. 217–135.

Bishop, G.F., Oldendick, R.W., Tuchfarber, A.J. & Bennett, S.E. (1979) Effects of opinion filtering and opinion floating: evidence from a secondary analysis. *Political Methodology*, **6**, pp. 293–309.

Blair, E., Sudman, S., Bradburn, N. & Stocking, C. (1977) How to ask questions about drinking and sex: response effects in measuring consumer behavior. *Journal of Marketing Research*, **14**, pp. 316–321.

Bortz, J. & Döring, N. (2003) *Forschungsmethoden und Evaluation für Sozialwissenschaftler*. Berlin, Heidelberg: Springer.

Bowling, A. (2005) Quantitative social science: the survey. In A. Bowling & S. Ebrahim (eds) *Handbook of Health Research Methods: Investigation, Measurement and Analysis*. New York: McGraw-Hill, pp. 190–214.

Brace, I. (2004) *Questionnaire Design. How to Plan, Structure and Write Survey Material for Effective Market Research*. London: Kogan Page.

Bradburn, N. & Miles, C. (1979) Vague quantifiers. *Public Opinion Quarterly*, **43**, 1, pp. 92–101.

Bradburn, N. & Sudman, S. (1979) *Improving Interview Method and Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Bradburn, N. & Sudman, S. (2003) The current status of questionnaire design. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (eds) *Measurement Errors in Surveys*. New York: Wiley, pp. 29–40.

Bradburn, N., Sudman, S. & Wansink, B. (2004) *Asking Questions. The Definitive Guide to Questionnaire Design – for Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco, CA: Jossey-Bass.

Brislin, R.W. (1986) The wording and translation of research instruments. In W.J. Lonner & J.W. Berry (eds) *Field Methods in Cross-Cultural Research*. Newbury Park, CA: Sage, pp. 137–164.
Cannell, C.F., Miller, P.V. & Oksenberg, L. (1981) Research on interviewing techniques. In S. Leinhardt (ed.) *Sociological Methodology*. San Francisco, CA: Jossey-Bass.
Cliff, N. (1959) Adverbs as multipliers. *Psychological Review*, **66**, pp. 27–44.
Coelho, P.S. & Esteves, S.P. (2007) The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement. *International Journal of Market Research*, **49**, 3, pp. 313–339.
Converse, J. & Presser, S. (1986) *Survey Questions. Handcrafting the Standard Questionnaire*. London: Sage.
Cox III, E.P. (1980) The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, **17**, pp. 407–422.
Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrica*, **16**, pp. 93–96.
Crowne, D. & Marlowe, D. (1960) A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, **24**, pp. 963–968.
Dawes, J. (2008) Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, **50**, 1, pp. 61–77.
deLeeuw, E. & deHeer, W. (2002) Trends in household survey nonresponse: a longitudinal and international comparison. In R. Groves, D. Dillman, J. Eltinge & R. Little (eds) *Survey Non-Response*. New York: Wiley, pp. 41–54.
DeVellis, R.F. (2003) *Scale Development. Theory and Application* (2nd edn). Thousand Oaks, CA: Sage.
Diekmann, A. (2003) *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Reinbeck bei Hamburg: Rowohlt.
Dillman, D. (2000) *Mail and Internet Surveys. The Tailored Design Method*. New York: John Wiley & Sons, Inc.
DiStefano, C. & Motl, R.W. (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, **13**, 3, pp. 440–464
Dörnyei, Z. (2003) *Questionnaires in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum.
Drennan, J. (2003) Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing*, **41**, 1, pp. 57–63.
Dudycha, A.L. & Carpenter, J.B. (1973) Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, **58**, pp. 116–121.
Edwards, A. (1957) *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden Press.
Eifermann, R.R. (1961) Negation: a linguistic variable. *Acta Psychologica*, **18**, pp. 258–273.
Fife-Schaw, C. (2006) Questionnaire design. In G.M. Breakwell, S. Hammond, C. Fife-Schaw, & J.A. Smith (eds) *Research Methods in Psychology* (3rd edn). Thousand Oaks, CA: Sage, pp. 210–231.
Fink, A. (2003) *How to Ask Survey Questions*. Thousand Oaks, CA: Sage.
Finn, R. (1972) Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, **32**, 2, pp. 255–265.
Fitzgerald, R. & Widdop, S. (eds) (2004) *European Social Survey Round 3. Measuring Social and Political Change in Europe*. Publishable final activity report. Available online at: http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=233:r3-far&catid=22:news&Itemid=48 (accessed 1 December 2008).

Foddy, W. (1993) *Constructing Questions for Interviews and Questionnaires. Theory and Practice in Social Research*. Cambridge, UK: Cambridge University Press.
Fowler, F.J. (1992) How unclear terms affect survey data. *Public Opinion Quarterly*, **56**, 2, pp. 218–231.
Fowler, F.J. (1995) *Improving Survey Questions. Design and Evaluation*. London: Sage.
Fowler, F.J. (2004) The case for more split-sample experiments in developing survey instruments. In S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer (eds) *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.
Garland, R. (1991) The mid-point on a rating scale: is it desirable? *Marketing Bulletin,* **2**, pp. 66–70.
Gaskell, G.D., O'Muircheartaigh, C.A. & Wright, D.B. (1994) Survey questions about the frequency of vaguely defined events: the effects of response alternative. *Public Opinion Quarterly*, **58**, 2, pp. 241–254.
Hippler, H.-J. & Schwarz, N. (1986) Not forbidding isn't allowing: the cognitive basis of the forbid–allow asymmetry. *Public Opinion Quarterly*, **50**, 1, pp. 87–96.
Hippler, H.-J., Schwarz, N. & Sudman, S. (1987) (eds) *Social Information Processing and Survey Methodology*. New York: Springer.
Holbrook, A., Cho, Y.I. & Johnson, T. (2006) The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly*, **70**, 4, pp. 565–595.
Holtgraves, T. (2004) Social desirability and self-reports: testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, **30**, 2, pp. 161–172.
Hunt, W.H., Sparkman, R.D. Jr & Wilcox, J.B. (1982) The pretest in survey research: issues and preliminary findings. *Journal of Marketing Research*, **19**, 2, pp. 269–273.
Jabine, T.B. (1987) Reporting chronic conditions in the National Health Interview Survey: a review of tendencies from evaluation studies and methodological test. *Vital and Health Statistics*, Series **2**, 105. Washington, DC: Government Printing Office.
Jobe, J. & Mingay, D. (1989) Cognitive research improves questionnaires. *American Journal of Public Health*, **79**, 8, pp. 1053–1055.
Kalton, G., Robert, J. & Holt, D. (1980) The effects of offering a middle response option with opinion questions. *The Statistician*, **29**, pp. 65–79.
Krosnick J.A. (1991) Response strategies for coping with the cognitive demands of attitude measurement in surveys. *Applied Cognitive Psychology,* **5**, 2, pp. 213–236.
Krumpal, I., Rauhut, H., Böhr, D. & Naumann, E. (2008) Wie wahrscheinlich ist 'wahrscheinlich'? *Methoden – Daten – Analysen*, **2**, 1, pp. 3–27.
Leite, W. & Beretvas, N. (2005) Validation of scores on the Marlowe–Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement*, **65**, 1, pp. 140–154.
Lenski, G.E. & Leggett, J.C. (1960) Caste, class and deference in the research interview. *American Journal of Sociology*, **65**, pp. 463–467.
Leverkus-Brüning, R. (1966) *Die Meinungslosen. Die Bedeutung der Restkategorie in der empirischen Sozialforschung*. Berlin: Duncker & Humbolt.
Liechtenstein, S. & Newman, J.R. (1967) Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, **9**, 10, pp. 563–564.
Litwin, M.S. (2003) *How to Assess and Interpret Survey Psychometrics* (2nd edn). Thousand Oaks, CA: Sage.
Martin, E. (2002) The effects of questionnaire design and reporting of detailed Hispanic origin in census 2000 mail questionnaires. *Public Opinion Quarterly*, **66**, 4, pp. 582–593.
Martin, M.O., Mullis, I.V.S. & Kennedy, A.M. (2007) *PIRLS 2006 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Masters, J.R. (1974) The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, **11**, 1, pp. 49–53.
Matell, M. & Jacoby, J. (1971) Is there an optimal number of alternatives for Likert scale items? Study I: reliability and validity. *Educational and Psychological Measurement*, **31**, pp. 657–74.
Mayer, H. (2002) *Interview und schriftliche Befragung. Entwicklung, Durchführung und Auswertung*. Munich: Oldenbourg Wissenschaftsverlag.
Mittelstaedt, R. (1971) Semantic properties of selected evaluative adjectives: other evidence. *Journal of Marketing Research*, **8**, 2, pp. 236–237.
Moorman, R.H. & Podsakoff, P.M. (1992) A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology*, **65**, pp. 131–149.
Mosier, C.I. (1941) A psychometric study of meaning. *Journal of Social Psychology*, **13**, pp. 123–140.
Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P. (2007) *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement (IEA).
Myers, J. & Warner, G. (1968) Semantic properties of selected evaluation adjectives. *Journal of Marketing Research*, **5**, pp. 409–412.
Nunnally, J.C. & Bernstein, I. (1994) *Psychometric Theory* (3rd edn). New York: McGraw-Hill.
O'Muircheartaigh, C., Gaskell, G. & Wright, D. (1993) Intensifiers in behavioral frequency questions. *Public Opinion Quarterly*, **57**, 4, pp. 552–565.
O'Muircheartaigh, C., Gaskell, G. & Wright, D. (1995) Weighing anchors: verbal and numeric labels for response scales. *Journal of Official Statistics*, **11**, 3, pp. 295–307.
O'Muircheartaigh, C., Krosnick, J. & Helic, A. (2000) Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. Unpublished manuscript. Retrieved 19 May 2008 from http://harrisschool.uchicago.edu/about/publications/working-papers/pdf/wp_01_3.pdf.
Oksenberg, L. & Cannell, C. (1977) Some factors underlying the validity of response in self report. *Bulletin of the International Statistical Institute*, **47**, pp. 325–346.
Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter.
Paulhus, D.L. (1984) Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, **46**, pp. 598–609.
Paulhus, D.L. (1991) Measurement and control of response bias. In J. Robinson, P. Shaver & L. Wrightsman (eds) *Measures of Personality and Social Psychological Attitudes*. San Diego, CA: Academic Press, pp. 17–59.
Paulhus, D.L. & Van Selst, M. (1990) The spheres of control scale: 10 years of research. *Personality and Individual Differences*, **11**, 10, pp. 1029–1036.
Porst, R. (2000) *Praxis der Umfrageforschung*. Stuttgart: Teubner.
Presser, S. & Blair, J. (1994) Survey pretesting: do different methods produce different results? *Sociological Methodology*, **24**, pp. 73–104.
Rammstedt, B. & Krebs, D. (2007). Response scale format and answering of personality scales. *European Journal of Psychological Assessment*, **23**, 1, pp. 32–38.
Rockwood, T., Sangster, R. & Dillman, D. (1997) The effects of response categories on questionnaire answers: context and mode effects. *Sociological Methods and Research*, **26**, 1, pp. 118–140.
Rodgers, W., Andrews, F. & Herzog, R. (1992) Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, **8**, 3, pp. 251–275.
Saris, W.E. & Gallhofer, I. (2007) Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods,* **1**, 1, pp. 29–43.

Saris, W.E., Satorra, A. & Coenders, G. (2004) A new approach for evaluating quality of measurement instruments. *Sociological Methodology*, **34**, 1, pp. 311–347.
Schaeffer, N.C. & Presser, S. (2006) The science of asking questions. *Annual Review of Sociology*, **29**, pp. 65–88.
Scholl, A. (2003) *Die Befragung. Sozialwissenschaftliche Methode und kommunikationswissenschaftliche Anwendung*. Konstanz: UVK Verlagsgesellschaft.
Schuman, H. & Presser, S. (1977) Question wording as an independent variable in survey analysis. *Sociological Methods and Research*, **6**, pp. 151–176.
Schuman, H. & Presser, S. (1978) Attitude measurement and the gun control paradox. *Public Opinion Quarterly*, **41**, pp. 427–439.
Schuman, H. & Presser, S. (1996) *Questions and Answers in Attitude Surveys*. London: Sage.
Schwarz, N. & Hippler, H. (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*. New York: Wiley Interscience, pp. 41–56.
Schwarz, N. & Sudman, S. (1992) *Context Effects in Social and Psychological Research*. New York: Springer.
Schwarz, N., Grayson, C.E. & Knäuper, B. (1998) Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research*, **10**, 2, pp. 177–183.
Schwarz, N., Hippler, H., Deutsch, B. & Strack, F. (1985) Response scales: effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, **49**, 3, pp. 388–395.
Schwarz, N., Knäuper, B., Hippler, H., Noelle-Neumann, E. & Clark, L. (1991) Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, **55**, pp. 570–582.
Seitz, W. (1977) *Persönlichkeitsbeurteilung durch Fragebogen. Eine Einführung in die diagnostische Praxis und ihre theoretischen Grundlagen für Psychologen, Pädagogen, Heilpädagogen und Mediziner*. Braunschweig: Westermann.
Simpson, R.H. (1944) The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech*, **21**, 3, pp. 328–330.
Smith, P.B. (2004) Acquiescent response bias as an aspect of cultural communications style. *Journal of Cross-Cultural Psychology*, **35**, pp. 50–61.
Smith, T.W. (2003) Developing comparable questions in cross-national surveys. In J.A. Harkness, F.J.R. van de Vijver & P.P. Mohler (eds) *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley Interscience, pp. 69–92.
Stocké, V. & Hunkler, C. (2007) Measures of desirability beliefs and their validity as indicators for socially desirable responding. *Field Methods*, **19**, 3, pp. 313–336.
Sudman, S. & Bradburn, N.M. (1974) *Response Effects in Surveys*. Chicago, IL: Aldine.
Tourangeau, R., Rips, L.J. & Rasinski, K. (2000) *The Psychology of Survey Response*. New York: Cambridge University Press.
Trometer, R. (1996) *Warum sind Befragte 'meinungslos'? Kognitive und kommunikative Prozesse im Interview. Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie der Universität zu Mannheim*. Mannheim: Universität Mannheim.
van der Zouwen, J. (2000) An assessment of the difficulty of questions used in the ISSP-questionnaires, the clarity of their wording, and the comparability of the responses. *ZA-Information,* **46**, pp. 96–114.
Vidali, J.J. (1975) Context effects on scaled evaluatory adjective meaning. *Journal of the Market Research Society*, **17**, 1, pp. 21–25.
Vikat, A., Speder, Z., Beets, G., Billari, F.C. & Buhler, C. (2007) Generations and gender survey (GGS): towards a better understanding of relationships and processes. *Demographic Research*, **17**, 14, pp. 389–440.

von der Linden, W.J. & Hambleton, R.K. (1997) *Handbook of Modern Item Response Theory*. New York: Springer.
Weems, G.H., Onwugbuzie, A.J. & Lustig, D. (2002) Profiles of respondents who respond inconsistently to positively and negatively worded items on rating scales. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Chattanooga, TN, 6–8 November.
White, P.C., Jennings, N.V., Renwick, A.R. & Barker, N.H.L. (2005) Questionnaires in ecology: a review of past use and recommendations for best practice. *Journal of Applied Ecology*, **42**, 3, pp. 421–430.
Wildt, A.R. & Mazis, M.B. (1978) Determinants of scale response: label versus position. *Journal of Marketing Research*, **15**, pp. 261–267.
Willis, G.B. (2005) *Cognitive Interviewing. A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.
Worcester, R. & Burns, T. (1975) A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society*, **17**, pp. 181–197.
Wright, B.D. & Masters, G.N. (1982) *Rating Scale Analysis*. Chicago, IL: MESA Press.

About the author

Dr Petra Lietz joined the Australian Council for Educational Research in February 2009 as a Senior Research Fellow in the National and International Surveys Program. She previously held the position of Professor of Quantitative Research Methods at Jacobs University Bremen, Germany, where she taught undergraduate and graduate courses in logic of comparative research, secondary data analysis, statistics and research design since 2003. Courses in research methods and design also formed part of her responsibilities at Central Queensland University in Rockhampton, Australia, where she was a Senior Lecturer in Research Methods and Associate Dean Research in the Faculty of Business from 1997–2000. Her research interests include survey research methodology and methodological issues in internationally comparative research, particularly questionnaire design. Her publications feature results of multivariate and multilevel analyses with a focus on factors influencing student achievement using data from the OECD Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), conducted by the International Association for the Evaluation of Educational Achievement (IEA). She is currently a member of the coordination team responsible for the PISA 2012 context questionnaire development.

Address correspondence to: Petra Lietz, Australian Council for Educational Research (ACER), Adelaide Office, Level 10, 60, Waymouth Street, Adelaide, S.A. 5000, Australia.

Email: lietz@acer.edu.au