1

Conceptualizing data quality: Respondent attributes, study architecture and institutional practices

Assessing the quality of data is a major endeavour in empirical social research. From our perspective, data quality is characterized by an absence of artefactual variation in observed measures. Screening survey data means searching for variation in observed responses that do not correspond with actual differences between respondents. We agree with Holbrook, Cho and Johnson (2006: 569) who argue that screening techniques are essential because survey researchers are 'far from being able to predict a priori when and for whom' comprehension or response mapping difficulties will occur; and these are only two of many sources of poor data quality.

We think of data quality as an umbrella concept that covers three main sources affecting the trustworthiness of any survey data: the study architecture, the institutional practices of the data collection agencies, and the respondent behaviours. Study architecture concerns elements in the survey design, such as the mode of data collection (e.g., computer-assisted telephone interviews, mailed questionnaires, internet surveys), the number of questions and the order in which they are asked, the number and format of the response options, and the complexity of the language employed. Institutional practices cover sources of error that are due to the research organization, such as the adequacy of interviewer training, appropriateness of the sampling design, and data entry monitoring procedures. Data quality is obviously also affected by respondent attributes, such as their verbal skills or their ability to retrieve the information requested. While we discuss these three sources of data quality separately, in practice they interact with each other in myriad ways. Thus, self-presentation issues on the part of the respondent, for example, play a larger role in face-to-face interviews than in internet surveys.

While quality of data is a ubiquitous research concern, we focus on assessing survey data quality. Our concern is with all aspects of data quality that jeopardize

the validity of comparative statistics. Group comparisons are compromised when the quality of data differs for the groups being compared or when the survey questions have different meanings for the groups being compared. If females are more meticulous than males in their survey responses, then gender differences that may emerge in subsequent analyses are suspect. If university-educated respondents can cope with double negative sentences better than those with less education, then educational differences on the distribution of such items are substantively ambiguous. In short, it is the *inequality* of data quality that matters most, since the logic of survey analysis is inherently comparative. If the quality of the data differs between the groups being compared, then the comparison is compromised.

We further restrict our attention to the underlying structure of responses to a set of statements on a particular topic or domain. This topic can be a concrete object such as the self, contentious issues such as national pride or regional identity, or nebulous concepts such as democracy. Respondents are typically asked to indicate the extent of their acceptance or rejection of each of the statements. Their responses are expected to mirror their viewpoints (or cognitive maps as they will be called here) on that topic or issue. Excluded from consideration in this book is the quality of socio-demographic and other factual information such as a person's age, income, education, or employment status.

In this chapter we first discuss the three sources of data quality, namely those attributable to the respondent, those arising from the study architecture, and those that emerge from inadequate quality control procedures of data collection agencies, including unethical practices. This is followed by a description of the nature and logic of our screening approach, which is anchored in scaling methods, especially multiple correspondence analysis and categorical principal component analysis. We conclude with a sketch of the key content of each of the subsequent chapters.

Conceptualizing response quality

1.1 We refer to sources of data quality that are due to respondents' characteristics, such as their response styles and impression management skills, as response quality. Response quality is embedded in the dynamics common to all human interactions as well as the specific ones that arise out of the peculiar features of survey protocol. Common features, as recognized by the medical field, for example, emerge from the fact that a survey 'is a social phenomenon that involves elaborate cognitive work by respondents' and 'is governed by social rules and norms' (McHorney and Fleishman, 2006: S206).

The act of obtaining survey data imposes a particular stylized form of human interaction, which gives rise to its specific dynamics. The parameters that govern the survey form of interaction are as follows:

- The contact and subsequent interaction is initiated by the interviewer, typically without the express desire of the respondent.
- It occurs between strangers, with one of the members not even physically present when the survey is conducted via mail, telephone, or the web.
- The interaction is a singular event with no anticipation of continuity, except in longitudinal and/or other panel surveys where the interaction is limited to a finite series of discrete events.
- Interactional reciprocity is violated; specifically, the interviewers are expected to ask questions while the respondents are expected to provide answers.
- The researcher selects the complexity level of the language and its grammatical style, which typically is a formal one.
- The response vocabulary through which the respondents must provide their responses is extremely sparse.

In short, surveys typically consist of short pulses of verbal interaction conducted between strangers on a topic of unknown relevance or interest to the respondent, often in an alien vocabulary and with control of the structure of the interaction vested in the researcher. What the respondent gets out of this unequal exchange is assurances of making a contribution to our knowledge base and a promise of confidentiality and anonymity, which may or may not be believed. Is it any wonder, then, that one meta-analysis of survey data estimated that over half the variance in social science measures is due to a combination of random (32%) and systematic (26%) measurement error, with even more error for abstract concepts such as attitudes (Cote and Buckley, 1987: 316)? Clearly, these stylistic survey features are consequential for response quality. Such disheartening findings nevertheless form the underpinnings and the rationale for this book, since data quality cannot be taken for granted and therefore we need tools by which it can be assessed.

Given the features of a survey described above, it is wisest to assume that responses will be of suboptimal quality. Simon (1957) introduced the term 'satisficing' to situations where humans do not strive to optimize outcomes. Krosnick (1991, 1999) recognized that the survey setting typically induces satisficing. His application is based on Tourangeau and his associates' (Tourangeau and Rasinski, 1988; Tourangeau, Rips and Rasinski, 2000) four-step cognitive process model for producing high-quality information: the respondent must (1) understand the question, (2) retrieve the relevant information, (3) synthesize the retrieved information into a summary judgement, and (4) choose a response option that most closely corresponds with the summary judgement. Satisficing can take place at any of these stages and simply means a less careful or thorough discharge of these tasks. Satisficing manifests itself in a variety of ways, such as choosing the first reasonable response offered, or employing only a subset of the response options provided. What all forms of satisficing have in common is that shortcuts are taken that permit the task to be discharged more quickly while still fulfilling the obligation to complete the task.

The task of responding to survey questions shares features with those of other literacy tasks that people face in their daily lives. The most important feature is

that responding to survey items may be cognitively challenging for some respondents. In particular, responding to lengthy items and those containing a negation may prove to be too demanding for many respondents – issues that Edwards (1957) noted more than half a century ago. Our guiding assumption is that the task of answering survey questions will be discharged quite differently among those who find this task daunting compared to those who find it to be relatively easy.

Faced with a difficult task, people often use one of three response strategies: (1) decline the task, (2) simplify the task, and (3) discharge the task, however poorly. All three strategies compromise the response quality. The first strategy, declining the task, manifests itself directly in outright refusal to participate in the study (unit non-response) or failing to respond to particular items by giving non-substantive responses such as 'don't know' or 'no opinion' (item non-response). Respondents who simplify the task frequently do this by favouring a subset of the available response options, such as the end-points of Likert-type response options, resulting in what is known as 'extreme response style'. Finally, those who accept the demanding task may just muddle their way through the survey questions, perhaps by agreeing with survey items regardless of the content, a pattern that is known as an acquiescent response tendency. Such respondents are also more likely to be susceptible to trivial aspects of the survey architecture, such as the order in which response options are presented. We concur with Krosnick (1991) that response quality depends on the difficulty of the task, the respondent's cognitive skill, and their motivation to participate in the survey. The elements of each of these are presented next.

Task difficulty, cognitive skills, and topic salience

The rigid structure of the interview protocol, in conjunction with the often alien vocabulary and restrictive response options, transforms the survey interaction into a task that can be cognitively challenging. Our guiding assumption is that the greater the task difficulty for a given respondent, the lower will be the quality of the responses given. Task characteristics that increase its difficulty are:

- numerous, polysyllabic, and/or infrequently used words;
- negative constructions (especially when containing the word 'not');
- retrospective questions;
- double-barrelled formulations (containing two referents but permitting only a single response);
- abstract referents.

Despite being well-known elements of task difficulty, it is surprising how often they are violated – even in well-known surveys such as the International Social Survey Program and the World Values Survey.

Attributes of the response options, such as their number and whether they are labelled, also contribute to task difficulty. Response options that are labelled can act to simplify the choices. Likewise, response burden increases with the number of response options. While minimizing the number of response options may simplify the task of answering a given question, it also diminishes the amount of the information obtained, compromising the quality of the data again. Formats that provide an odd number of response options are generally considered superior to even-numbered ones. This may be because an odd number of response options, such as a five- or 11-point scale, provides a mid-point that acts as a simplifying anchor for some respondents.

Whether the survey task is difficult is also a matter of task familiarity. The format of survey questions is similar to that of multiple choice questions on tests and to application forms for a variety of services. Respondents in non-manual occupations (and those with higher educational qualifications) are more exposed to such forms than their counterparts in manual occupations (and/or with less formal education). Public opinion surveys are also more common in economically developed countries, and so the response quality is likely to be higher in these countries than in developing countries.

Van de Vijver and Poortinga (1997: 33) point out that 'almost without exception the effects of bias will systematically favor the cultural group from where the instrument originates'. From this we formulate the cultural distance bias hypothesis: the greater the cultural distance between the origin of a survey instrument and the groups being investigated, the more compromised the data quality and comparability is likely to be. One source of such bias is the increased mismatch between the respondent's and researcher's 'grammar' (Holbrook, Cho and Johnson, 2006: 569). Task difficulty provides another possible rationale for the cultural distance bias hypothesis, namely that the greater the cultural distance, the more difficult is the task of responding to surveys. The solution for such respondents is to simplify the task, perhaps in ways incongruent with the researcher's assumptions.

Whether a task is difficult depends not only on the attributes of the task but also on the cognitive competencies and knowledge of the respondent, to which we turn next. Cognitive skills are closely tied to education (Ceci, 1991). For example, the research of Smith et al. (2003) suggests that elementary school children do not have the cognitive sophistication to handle either a zero-to-ten or a thermometer response format – formats that generally have solid measurement properties among adults (Alwin, 1997). Likewise, understanding that disagreement with a negative assertion is equivalent to agreement with a positively formulated one remains problematic even for some high school students (Marsh, 1986; Thiessen, 2010).

Finally, we assume that respondents pay greater heed to tasks on topics that interest them. Generally these are also the ones on which they possess more information and consequently also the issues for which providing a valid response is easier. Our approach to response quality shares certain features with

that of Krosnick's (1991, 1999) satisficing theory, which emphasizes the cognitive demands required to provide high-quality responses. For Krosnick, the probability of taking shortcuts in any of the four cognitive steps discussed previously decreases with cognitive ability and motivation, but increases with task difficulty. We agree that response optimizing is least prevalent among those with least interest or motivation to participate in a survey.

Normative demands and impression management

Surveys share additional features with other forms of verbal communication. First, the form of survey interaction is prototypically dyadic: an interviewer/researcher and a respondent in real or virtual interaction with each other. In all dyadic interactions, the members hold at least three images of each other that can profoundly affect the content of the viewpoints the respondent expresses: the image of oneself, the image of the other, and the image one would like the other to have of oneself. It is especially the latter image that can jeopardize data quality. Skilful interactions require one to be cognizant not only about oneself and the other, but also about how one appears to the other. Hence, the responses given are best conceived of as an amalgam of what respondents believe to be true, what they believe to be acceptable to the researcher or interviewer, and what respondents believe will make a good impression of themselves. Such impression management dynamics are conceptualized in the methodological literature as social desirability.

Second, we ordinarily present ourselves as being more consistent than we actually are. This is exemplified by comparing the determinants of actual voting in elections with those of reported voting. Typically the associations between various civic attitudes and self-reported voting behaviour are stronger than with actual (validated) voting behaviour (Silver, Anderson and Abramson, 1986). That is, our reported behaviours are more consistent with our beliefs than are our actual behaviours. If respondents initially report that they intended to vote, then they subsequently will be more likely to report that they voted even when they did not. Likewise, respondents who report that it is one's civic duty to vote are more likely to report that they voted when they did not compared to those who did not think it was one's civic duty.

Third, the normative structure places demands on the participants, the salience and extent of which depend on one's social location in society. The social location of some respondents may place particular pressure on them to vote, or not to smoke, for example. These demand characteristics result in tendencies to provide responses that are incongruent with the positions actually held. The existing methodological literature also treats these pressures primarily under the rubric of social desirability, but we prefer the broader term of 'impression management'. Van de Vijver and Poortinga (1997: 34) remind us that '[n]orms about appropriate conduct differ across cultural groups and the social desirability expressed in assessment will vary accordingly'.

It is precisely the existence of normative demands that required modifications to classical measurement theory. This theory provided a rather simple measurement model, whereby any individual's observed score (y_i) is decomposed into two parts: true (τ_i) and error (ε_i); that is, $y_i = \tau_i + \varepsilon_i$. While this formulation is enticingly simple, the problem emerges with the usual assumptions made when applied to a distribution. If one assumes that the error is uncorrelated with the true score, then the observed (total) variance can be decomposed into true and error variance: $\text{Var}_y = \text{Var}_\tau + \text{Var}_\varepsilon$.

This decomposition is at the heart of reliability coefficients, which express reliability as the ratio of the true variance to the total variance. Of course, frequently the uncorrelated error assumption is untenable. One example should suffice: virtually all voting measurement error consists of over-reporting (Bernstein, Chadha and Montjoy, 2001; Silver, Anderson and Abramson, 1986). That is, if a person voted, then there is virtually no error, but if a person did not vote, then there is a high likelihood of (systematic) measurement error. The reason for this is that voting is not normatively neutral. Simply stated, the stronger the normative pressure, the greater is the systematic measurement error. Since societal norms surround most of the issues that typically form the content of survey questionnaires, it follows that the assumption of random measurement error is seldom justified.

Normative pressures are unequally socially distributed, having greater force for some. Returning to the voting example, Bernstein, Chadha and Montjoy (2001) argue that the normative pressure to vote is greater for the more educated and politically engaged. For this reason, these respondents are more likely to claim to have voted when they did not than their counterparts, casting considerable doubt on the estimated strengths of the relationships between education and political interest on the one hand, and voting on the other. Likewise, younger children are under greater pressure not to smoke than older ones. Hence, younger children who smoke are more likely to deny smoking than older ones (Griesler et al., 2008).

Normative demands are not the only source of systematic bias. Podsakoff et al. (2003) summarize 20 potential sources just of systematic bias. While not all the biases they list are relevant to all survey research, their literature review sensitizes us to the complex array of factors that can result in systematic biases. The authors conclude that 'methods biases are likely to be particularly powerful in studies in which the data for both the predictor and criterion variable are obtained from the same person in the same measurement context using the same item context and similar item characteristics' (Podsakoff et al., 2003: 885). That, unfortunately, describes the typical survey.

Campbell and Fiske's (1959) documentation of high proportions of common method variance led to a revision of classical measurement theory to incorporate the likelihood of reliable but invalid method variance. In structural equation modelling language (Alwin, 1997), an observed score can be decomposed

into three unobserved components: $y_{ij} = \lambda_i \tau_i + \lambda_j \eta_j + \varepsilon_{ij}$, where y_{ij} measures the *i*th trait by the *j*th method, τ_i is the *i*th trait, and η_j the *j*th method factor. The λ_i can be considered the validity coefficients, the λ_j as the invalidity coefficients, and ε_{ij} is the random error. This formulation makes explicit that some of the reliable variance is actually invalid, that is, induced by the method of obtaining the information.

Systematic error is crucially important since it provides an omnipresent alternative explanation to any substantive interpretation of a documented relationship. When a substantive and an artefactual interpretation collide, by the principle of parsimony (Occam's razor) the artefactual one must win, since it is the simpler one. As a direct consequence of this, solid research should first assess whether any of the findings are artefactual. A variety of individual response tendencies, defined collectively as the tendency to disproportionately favour certain responses, has been the subject of much methodological research, since they could be a major source of artefactual findings. Response tendencies emerge out of the response options that are provided, which is part of the questionnaire architecture, to which we turn next.

Study architecture

1.2 Data quality is compromised when the responses selected are not in agreement with the cognitions held. One reason for such disparities has to do with the response options provided. Most frequently, the respondent's task in a survey is to select one response from a set of response options. The set of response options is collectively referred to as the response format. These formats range from simple yes–no choices to 0–100 thermometer analogues and, more recently, visual analogue scales in web surveys (Reips and Funke, 2008). One of the most popular response formats is some variant of the Likert scale, where response options typically vary from 'strongly agree' at one end to 'strongly disagree' at the other. Some response formats explicitly provide for a non-substantive response, such as 'don't know', 'no opinion' or 'uncertain'.

Data comparability is compromised whenever respondents with identical viewpoints employ different response options to express them. The survey literature makes two conclusions about response options abundantly clear. The first is that some responses are more popular than others. For example, a response of 'true' is more likely to be selected than a response of 'false' (Cronbach, 1950). Likewise, in the thermometer response format, responses divisible by 10 are far more favoured than other responses (Kroh, 2007). Second, respondents themselves differ in the response options they favour. Some respondents eschew extreme responses; others favour non-substantive responses such as 'don't know'.

Response options and response tendencies

Response tendencies represent reproducible or systematic variation that remains after the variation due to item content has been removed. For example, Oskamp (1977: 37) defines response sets as 'systematic ways of answering which are not directly related to the question content, but which represent typical behavioural characteristics of the respondents'. Generally, response style produces systematic measurement error rather than random measurement error. Being partial to certain response options has spawned a huge literature under the rubric of response styles and response sets. Four of them are considered to be particularly consequential. These are:

Acquiescence response style (ARS) is the tendency to provide a positive response, such as *yes* or *agree*, to any statement, regardless of its content. ARS is premised on the observation that endorsing survey items is more common than rejecting them. That is, responses of 'yes', 'true', and various shades of 'agree' are more common than their counterparts of 'no', 'false', and levels of 'disagree'. Although its logical opposite, the tendency to disagree, has also been identified as a possible response style, it has received little empirical attention. One reason is that disagreement is less common than agreement. A further reason is that empirically these two response tendencies are so strongly inversely related that it has not been fruitful to differentiate between them (Baumgartner and Steenkamp, 2001). Finally, as will be detailed in the next chapter, the theoretical underpinnings for ARS are more solid than for its opposite.

It is easiest to differentiate ARS from genuine substantive agreement when an issue is paired with its semantic opposite. Respondents who agree with a statement and its semantic opposite present themselves as logically inconsistent and this inconsistency is generally attributed to either ARS or a failure to pay attention to the content of the question – that is, to a failure to optimize. This is the only response tendency that can materialize with dichotomous response options.

Extreme response style (ERS) refers to the tendency to choose the most extreme response options available (such as 'strongly agree' and 'strongly disagree'). This response tendency can manifest itself only on response formats that have at least four available response options that distinguish intensity of viewpoints. Examples could be the choice of 'always' and 'never' in frequency assessments, or 'strongly agree' and 'strongly disagree' in Likert formats.

Mid-point responding (MPR) consists of selecting a neutral response such as 'neither agree nor disagree' or 'uncertain'. Not much research has been conducted on this response tendency, compared to ARS and ERS. Theoretically, its importance is related to the fact that it is a safe response, requiring little justification. It shares this aspect with non-substantive responses such as 'don't know' and 'no opinion' without the disadvantages of having to admit lack of knowledge or appearing uncooperative. Sometimes the use of the neutral response constitutes a 'safe' form of impression management whereby one can

seem to offer an opinion when one fails to have one (Blasius and Thiessen, 2001b; Thiessen and Blasius, 1998).

Limited response differentiation (LRD) arises when respondents tend to select a narrower range of responses out of those provided to them. Typically it is measured by the individual's standard deviation across a battery of items. LRD differs from the other response tendencies in that it is less specific, that is, it does not focus attention on a specific response, but rather on a more global lack of discrimination in responses across items. Like the other response tendencies, it can be viewed as a respondent's strategy to simplify the task at hand.

Other response styles have been identified, such as random or arbitrary response style (Baumgartner and Steenkamp, 2001; Watkins and Cheung, 1995). To the extent that they are actually response styles, what they have in common is either an effort to simplify the task or impression management.

Institutional quality control practices

1.3 Collecting quality survey data requires inordinate financial, technical, and human resources. For this reason such data gathering is usually conducted by public and private institutions specializing in the implementation of survey designs. All of the publicly available data sets we analyse in subsequent chapters were commonly designed by groups of experts but contracted out for implementation to different national data collection organizations. Hence, national variation in the quality of the data is to be expected.

However, data collection agencies operate under financial constraints: private agencies must make a profit to survive and public agencies must operate within their given budgets. This means that a tension between quality and cost arises in all aspects of the production of survey data. This tension leads us to postulate that the satisficing principle applies to data collection agencies and interviewers as much as it does to individual respondents. That is, organizations may collect 'good enough' data rather than data of optimal quality. But what is good enough? To give an example from face-to-face interviews, several methods can be used to draw a sample. The best – and most expensive – is to draw a random sample from a list provided by the statistical office of a country (city) that contains the names and addresses of all people in the relevant population. The survey organization then sends an official letter to the randomly selected target persons to explain the purpose of the survey and its importance, and to allay fears the respondents may have about the legitimacy of the survey. Interviewers are subsequently given the addresses and required to conduct the interviews specifically with the target persons. If the target person is unavailable at that time, they are not permitted to substitute another member of that household (or a neighbouring one). A much worse data collection method – but a relatively cheap one – is known as the random route: interviewers must first select a household according to a fixed rule, for example, every fifth household. Then they must select a

target person from the previously selected household, such as the person whose birthday was most recent. In this design, refusals are likely to be high, since there was no prior contact explaining the study's purpose, importance, or legitimacy. Further, the random route method is difficult (and therefore costly) for the survey institute to monitor: which is the fifth household, and whose birthday really was the most recent? If the interviewer made a counting error, but the interview is well done, there is no strong reason to exclude the interview.

As Fowler (2002) notes, evidence shows that when interviews are not tape-recorded, interviewers are less likely to follow the prescribed protocol, and actually tend to become less careful over time. But since monitoring is expensive rather than productive, it ultimately becomes more important to the institute that an interview was successfully completed than that it was completed by the right respondent or with the correct protocol.

1.4 Data screening methodology

The screening methods we favour, namely multiple correspondence analysis (MCA) and categorical principal component analysis (CatPCA), are part of a family of techniques known as scaling methods that are described in Chapter 3. However, our use of these scaling techniques is decidedly not for the purpose of developing scales. Rather, it is to visualize the response structure within what we call the respondents' cognitive maps of the items in the domain of interest. Typically, these visualizations are represented in a two-dimensional space (or a series of two-dimensional representations of higher-order dimensions). Each distinct response to every item included in the analysis is represented geometrically in these maps. As will be illustrated in our substantive analyses, the location of each response category relative to that of all other response categories from all items provides a rich set of clues about the quality of the data being analysed.

MCA and CatPCA make relatively few assumptions, compared to principal component analysis (PCA) and structural equation modelling (SEM). They do not assume that the data is metric, and MCA does not even assume that the responses are ordinal. Additionally, there is no need to start with a theoretical model of the structure of the data and its substantive and non-substantive components. Rather, attention is focused on the geometric location of both responses and respondents in a low-dimensional map. Screening the data consists of scrutinizing these maps for unexpected or puzzling locations of each response to every item in that map. Since there are no prior models, there is no need for fit indices and there is no trial-and-error procedure to come up with the best-fitting model. We are engaged simply in a search for anomalies, and in each of our chapters we discovered different anomalies. We need no model to predict these, and indeed some of them (such as the ones uncovered in Chapter 4 on institutional practices) rather surprised us.

To minimize nonsensical conclusions, standard practice should include data screening techniques of the types that we exemplify in subsequent chapters. What we do in our book can be thought of as the step prior to assessing configural invariance. Scholarly research using survey data is typically comparative: responses to one or more items on a given domain of interest by one group are compared to those given by another group. Comparative research should start with establishing the equivalence of the response structures to a set of items. The minimum necessary condition for obtaining construct equivalence is to show that the cognitive maps produced by the item battery have a similar underlying structure in each group. Issues of response bias are not the same as either reliability or validity. In response bias, the issue is whether the identical response to a survey item by different respondents has the same meaning. In our work, the meaning is inferred from the cognitive maps of groups of respondents who differ in some systematic way (such as cognitive ability, culture, educational attainment and race). If the cognitive maps are not similar, then the identical response is assumed to differ in meaning between the groups.

We apply MCA and CatPCA to the series of statements as a way to describe the structure underlying the overt responses. Analogous to PCA, these techniques locate each of the responses (as well as each of the respondents) in a lower-dimensional space, where the first dimension accounts for the greatest amount of the variation in the responses and each successive dimension accounts for decreasing proportions of such variation. Our guiding assumption is that if the data are of high quality, then the dimensions that reflect coherent substantive patterns of endorsement or rejection are also the ones that account for the greatest proportion of variance. If, on the other hand, the primary dimensions reflect methodological artefacts, or are not interpretable substantively, then we conclude that the data are of low quality.

Our assessment of data quality is stronger when the data have the following characteristics. First, the data set includes multiple statements on the focal domain. Generally, the more statements there are, the easier it is to assess their quality. Second, some of the statements should be formulated in reverse polarity to that of the others, which will allow one to assess whether respondents are cognizant of the direction of the items. Third, the item set should be somewhat heterogeneous; not all items need be manifestations of a single substantive concept.

1.5 Chapter outline

In this chapter we have provided an overview of the sources of data quality. Chapter 2 gives an overview of the empirical literature that documents the existence and determinants of data quality, with a heavy emphasis on factors affecting response quality. A variety of methods have been employed to detect and control for these sources of response quality, each with

its own set of advantages and limitations. The empirical review reinforces our view that what response styles have in common is that they represent task simplification mechanisms. There is little evidence that different response styles are basically distinct methodological artefacts. It also shows that cognitive competencies and their manifestations such as educational attainment have particularly pervasive effects on all aspects of response quality. Finally, it highlights some special problems in conducting cross-national (and especially cross-cultural) research.

Chapter 3 gives a brief overview of the methodological and statistical features of MCA and CatPCA, and their relationship to PCA, which helps to explain why these are our preferred methods for screening data. To exemplify the similarities and differences of the methods, we use the Australian data from the 2003 International Social Survey Program (ISSP) on national identity and pride. This overview assumes readers have some familiarity with both matrix algebra and the logic of multivariate statistical procedures. However, the subsequent chapters should be comprehensible to readers who have a basic knowledge of statistical analyses.

Our first exercise in data screening is given in Chapter 4 in which three different data sets are analysed and in which a different source of dirty data was implicated. The examples show how the anomalies first presented themselves and how we eventually found the source of the anomalies. In all three data sets, the problem was not the respondent but the survey organization and its staff. The 2005–2008 World Values Survey (WVS) was employed for two of the analyses in this chapter. The first shows a high probability that some interviewers were prompting the respondents in different ways to produce very simple and often-repeated response combinations that differed by country. The second analysis shows that the survey organizations in some countries engaged in unethical behaviours by manufacturing some of their data. Specifically, we document that some of the data in some countries were obtained through the simple expedient of basically a copy-and-paste procedure. To mask this practice, a few fields were altered here and there so that automated record comparison software would fail to detect the presence of duplicated data. Secondly, we show how to detect (partly) faked interviews using data from our own survey based on sociological knowledge of stereotypes. The last example, using data from the 2002 European Social Survey (ESS), shows that even fairly small errors in data entry (in the sense that they represented only a tiny proportion of all the data) were nevertheless detectable by applying MCA.

We use the ESS 2006 in Chapter 5 to provide an example of data we consider to be of sufficiently high quality in each of the participating countries to warrant cross-national comparisons. We conclude that the relatively high quality and comparability of the data used in this chapter is due to the low cognitive demands made on respondents. The construction of the questions was extremely simple on a topic for which it is reasonable to assume that respondents had direct knowledge, since the questions involved how often they had various

feelings. However, the example also shows that reliance on traditional criteria for determining the number of interpretable factors and the reliability of the scale can be quite misleading. Furthermore, we show that the common practice of rotating the PCA solutions that contain both negatively and positively formulated items can lead to unwarranted conclusions. Specifically, it suggests that the rotation capitalizes on distributional features to create two unipolar factors where one bipolar factor arguably is more parsimonious and theoretically more defensible. An additional purpose of this chapter was to show the similarities and differences between the PCA, MCA, and CatPCA solutions. It shows that when the data are of high quality, essentially similar solutions are obtained by all three methods.

The purpose of Chapter 6 is to show the adverse effect on response quality of complicated item construction in measures of political efficacy and trust. On the basis of the 1984 Canadian National Election Study (CNES), we show that the location of response options to complex questions is theoretically more ambiguous than for questions that were simply constructed. By obtaining separate cognitive maps for respondents with above- and below-average political interest, we document that question complexity had a substantially greater impact on the less interested respondents.

Chapter 7 uses the same data to exemplify respondent fatigue effects, capitalizing on the feature that in the early part of the interview respondents were asked about their views of federal politics and politicians, while in the latter part of the interview the same questions were asked regarding provincial politics and politicians. In this chapter we also develop the dirty data index (DDI), which is a standardized measure for the quality of a set of ordered categorical data. The DDI clearly establishes that the data quality is markedly higher for the federal than the provincial items. Since interest in provincial politics was only moderately lower than that for federal politics, we conclude that the lower data quality is due to respondent fatigue.

The links between cognitive competency (and its inherent connection to task difficulty) and task simplification dynamics that jeopardize data quality are explored in Chapter 8. For this purpose we use the Programme for International Student Assessment (PISA) data, which focused on reading and mathematics achievement in 2000 and 2003, respectively. Our analyses show consistent patterns in both data sets documenting that the lower the achievement level, (1) the higher the likelihood of item non-response, (2) the higher the probability of limited response differentiation, and (3) the lower the likelihood of making distinctions between logically distinct study habits. The results suggest that the cognitive maps become more complex as cognitive competence increases.