

Chapter 3

Methodological considerations

This chapter addresses potential measurement error in trust survey data, focusing on the way survey design can either exacerbate or mitigate it. For a range of issues (question wording, response formats, survey context, survey mode, cross-cultural response styles), evidence on the key methodological challenges for data quality and key practical messages for survey design are highlighted. The chapter concludes by pointing out areas where additional methodological research will be needed.

3.1. Introduction

This chapter addresses the potential for measurement error in survey data on trust and the way it interacts with – and can be influenced by – the survey design. Following a brief overview of the different types of response biases that respondents can exhibit and that can cause measurement error, the chapter discusses various methodological aspects of the design of questions and surveys that impact on these biases.

The chapter specifically discusses the methodological issues that are of particular importance for measuring trust. These include question wording (Section 3.3), response formats (Section 3.4), the survey context (Section 3.5), survey mode (Section 3.6) and (cross-cultural) response styles (Section 3.7). Each of these sections presents evidence on the key methodological challenges for data quality and highlights key messages for survey design. Wherever possible, the evidence is drawn from trust-specific studies. However, as the methodological literature on trust questions is quite meagre, broader evidence from self-reported measures is relied on heavily. In this context, the 2013 OECD Guidelines on the Measurement of Subjective Well-being provide a useful reference point for a much broader discussion of methodological considerations in measuring intangible concepts. The chapter concludes with directions for further research.

3.2. Measurement error

It is important to recognise that all measures, even objective ones, exhibit some degree of error. Hence, the goal is not to select a perfect (and probably non-existent) measure but rather one that is “good enough” to distinguish meaningful patterns, such as changes in trust over time and differences between population subgroups, from noise in the data. As trust items are relatively sensitive to varying survey conditions and to how questions are framed, advice on their measurement needs to be more specific than is the case for some more “objective” indicators, such as educational attainment. Arguably, however, this sensitivity also exists in many other self-reported survey measures that are already being collected, such as subjective well-being measures or income.

Measurement error is the extent to which survey measures reflect concepts other than those intended by the surveyor (OECD, 2013). This error can be either systematic, leading to a bias in the data that is consistent in some way, and that might result in, for example, values that are consistently higher or lower than might be expected, or random, i.e. varying in an unpredictable manner (Maggino, 2009). The risk of error is essentially the product of a complex interaction between methodological factors (such as the cognitive demands made by certain questions, or the contextual features of a survey that might influence responses), respondent factors (such as motivation, fatigue and memory), and the construct of interest itself (such as how interesting or relevant respondents find the survey).

In order to answer any survey question, respondents are assumed to go through several cognitive steps, which may be performed either sequentially or in parallel. These steps include understanding the question, recalling information from memory, forming a judgement,

formatting the judgement to fit the response alternatives, and editing the final answer before delivering it to the surveyor (Sudman, Bradburn and Schwarz, 1996). It is important to understand this question-answer process not as a simple robotic task, but as part of a social interaction process between respondent, interviewer, question design and the survey context.

Aspects of survey design and context can either cause or exacerbate measurement error. Respondent failures in memory, motivation, communication or knowledge, which all can lead to respondent error in self-reported measures, are often associated with an increased risk of response biases and the use of response heuristics (Bradburn, Sudman and Wansink, 2004). Response biases refer to particular patterns or distortions in how individuals or groups of individuals respond to questions, while response heuristics refer to (often sub-conscious) mental shortcuts that respondents rely on to choose their answers (OECD, 2013). Drawing on the classifications of Podsakoff et al. (2003), Table 3.1 provides an overview of the response biases and heuristics commonly associated with all self-reported measures (OECD, 2013). Some, but not all, can apply to trust measures. The following sections consider the various methodological features of survey design that can lead to these most relevant response biases.

Table 3.1. **Overview of response biases and heuristics**

Response bias or heuristic	Exhibited response pattern
Acquiescence or yea-saying	A tendency to agree with, or respond positively to, survey items regardless of their content.
Nay-saying	A tendency to disagree with, or respond negatively to, survey items regardless of their content.
Extreme responding	A tendency to use response categories towards the ends of a response scale/the most extreme response category.
Moderate responding	A tendency to use responses towards the middle of the response scale/the most moderate response category.
No-opinion responding	A tendency to select the response category that is most neutral in its meaning (e.g. neither agree nor disagree).
Random responding	A tendency to respond randomly, rather than meaningfully.
Digit preferences	On numerical response formats, a tendency to prefer using some numbers more than others.
Primacy effects	A tendency to select one of the first response categories presented on a list.
Recency effects	A tendency to select one of the last response categories presented on a list.
Socially desirable responding	Conscious or subconscious tendency to select response options more likely to conform with social norms or present the respondent in a good light.
Demand characteristics	A reaction to subtle cues that might reflect the surveyor's beliefs about how they should respond and/or their own beliefs about the purpose of the survey (e.g. "leading questions", where the tone or phrasing of the question suggests to respondents that particular answers should be favoured).
Consistency motif or bias	A tendency for respondents to try and ensure consistency between responses (e.g. consistency between a question about attitudes towards smoking and a question about cigarette purchasing habits).
Priming effects	Where the survey context (e.g. question order; survey source) influences how questions are understood or makes certain information more easily accessible to respondents.

Source: OECD (2013), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264191655-en>.

StatLink  <http://dx.doi.org/10.1787/888933584146>

3.3. Question wording

Evaluations of the possible effects of question construction are central to guaranteeing the comparability, internal consistency and test-retest reliability of seemingly similar yet slightly different survey items. Questions that are easily understood and not ambiguous and that do not pose an unnecessary burden on respondents will increase the validity of responses and at the same time depress error variability (OECD, 2013).

For trust measures, question wording encompasses aspects of question comprehension, translatability across languages and societal subgroups, and changes in wording. This section reviews the challenges that can arise from these aspects, with a particular focus on the effect of question wording (as most available studies focus on this aspect).

The issues

First, if we want to compare self-reported items of trust across survey participants and countries and reduce variability due to measurement error, it is essential that respondents comprehend and interpret questions in a similar and unambiguous way. This also includes translatability across languages between countries and across different socio-economic and demographic subgroups within a society. For example, survey comparability is compromised if a certain wording of trust questions evokes different connotations for old vs. young people or does not conceptually exist in certain cultures.

Second, an important question in survey design, particularly for attitudinal questions like trust measures, relates to whether micro-changes in question wording significantly influence results. On the one hand, it would be worrying if changes in question wording do not produce different results, as this would imply that respondents do not understand question nuances well and do not cognitively process the concepts that the different questions are trying to distinguish between. On the other hand, it can be problematic if a slightly different wording of items that have been designed to measure the same underlying construct (and are often used interchangeably in comparative research) actually leads respondents to interpret the questions in dissimilar ways.

The evidence

A common way to evaluate question comprehension (apart from cognitive testing, the results of which are rarely published) is to look at an item's response latencies, or the time taken to process a question and deliver an answer. However, it is not absolutely clear whether short response latencies for an item indicate whether it was well understood or whether the respondent answered randomly. Unfortunately, in either case no study has so far considered response speed with regard to trust questions. The best evidence that respondents have understood question meaning and provided meaningful answers is demonstrated by strong correlations between the measures themselves and real-world outcomes as well as other non-survey indicators (OECD, 2013). As highlighted in the previous chapter, considerable support does exist for the overall validity of trust measures, particularly with regards to interpersonal trust.

We now turn to the issue of question wording. Two central aspects will be considered in the following – how similar question wording has to be to capture the same underlying concept, as well as how specific question wording has to be to conclude that it indeed taps into trust.

When it comes to evaluative measures of interpersonal trust, most items that can be found across surveys and time do not very greatly depart from the original Rosenberg question: “Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?” Nevertheless, a few surveys, sometimes routinely and sometimes as part of intentional methodological research, have featured different versions of the original trust question. Although not all studies are split-sample experiments that would allow for the definite attribution of different response distributions to changes in question phrasing, the available evidence suggests that responses to interpersonal trust measures are quite sensitive to wording changes.

For example, Smith (1997) capitalises on the fact that the American General Social Survey (GSS), which has featured the Rosenberg scale as a core battery item since 1972, has additionally included variants of the standard question over the years. An examination of

these shows quite different response patterns between versions. In a 1983 GSS experiment, 57% of respondents answered yes to the question “do you think that most people can be trusted?”, while only 36.5% indicated that *most people can be trusted* when the item was phrased “Some people say that most people can be trusted. Others say you can’t be too careful in your dealing(s) with people. How do you feel about it?” Both items offered a dichotomous answering scale. A drop like this of 20.5% of trusting respondents between question versions is quite extraordinary. As Smith himself notes, this difference might be driven not only by question wording effects but also by response scale and order effects, as the items feature slightly different response options (a simple *yes/no* vs. a more explicit *most people can be trusted/you can’t be too careful*). Further, both items were preceded by questions on quite different other topics. Response scale type and question order within the survey context are potentially large sources of error variability, matters which are discussed in Sections 3.4 and 3.5. But even just considering the known impacts of question construction documented in the literature on attitudinal questions more generally, a different response distribution and the direction of difference between the two questions intuitively makes sense: on the one hand, *do you think that most people can be trusted?* is an unbalanced question that only specifies one direction of trust, instead of also spelling out the alternative *or do you think that most people cannot be trusted?* For other attitudinal questions beyond trust, which also commonly refer to intangible and ambiguous concepts, such balance effects have often been associated with encouraging acquiescence (Schuman and Presser, 1996; Peabody, 1961). On the other hand, the second question version, just like the standard Rosenberg question, includes a reference to *being careful*, which one could argue is not quite the same as distrust and thus introduces a different underlying concept to the respondent.

It has been recognised in other places that the concept of caution in the trust question might be problematic, since being careful could carry quite different connotations for different population subgroups. Hooghe and Ressken (2008) write that “carefulness might imply something else for someone who is weak and vulnerable, compared to an athletic, bright and well-off person.”

Empirical support for this hypothesis has been offered by at least two studies: Soroka, Helliwell and Johnston (2007) examine four different versions of interpersonal trust questions included in the 2000/2001 Equality, Security and Community (ESC) survey carried out in Canada,¹ two of which feature a “caution rider”. First, the authors also find evidence of the above-mentioned balance and acquiescence effects, depending on whether the question is offering a full balance of trust/non-trust alternatives. Second and more importantly, the authors assert that “saying that you cannot be too careful in dealing with people is not the same as saying that people in general cannot be trusted. The first represents a cautious disposition, while the latter simply represents the reverse of the question ‘would you say that most people can be trusted?’” (ibid., p. 114) They further find that while women are less trusting than men when the standard (balanced) trust question is used, they are more trusting than men when a question without the *cannot be too careful* rider is used. It therefore may be cautiousness, rather than trustworthiness, that drives gender differences in the Rosenberg trust question.

The second piece of supporting evidence comes from two methodological experiments using split samples that the Office for National Statistics (ONS) in the United Kingdom carried out in co-operation with the OECD in October 2015 and May 2016 during its standard Opinions Survey. As each split sample in this experiment featured a relatively small sample size of 500, one cannot interpret the following results as absolutely definite – they are not


statistically significant unless explicitly stated so. Nevertheless, they confirm the intuition of a caution rider effect on certain groups that came out elsewhere. In the first experiment (see Table 3.2), respondents were offered either the standard Rosenberg question with a caution rider: “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?” (with a dichotomous answering choice) or the more neutral 11-point scale European Social Survey version: “On a scale where 0 is not at all and 10 is completely, in general how much do you think people can be trusted?”

Table 3.2. **Comparison of interpersonal trust questions by gender and age groups**

Experimental round	Question wording	Response scale	Population				
			All	Men	Women	16-44 years	Above 45 years
October 2015	“Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”	Dichotomous	32.1	33.6	30.7	31.8	32.3
	“On a scale where 0 is not at all and 10 is completely, in general how much do you think people can be trusted?”	11-point scale	38.6	36.8	40.4	33.7*	43.1*
May 2016	“Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”	Dichotomous	35.6	37.6	33.7	31.4*	40*
	“Generally speaking, would you say that most people can be trusted, or that you cannot be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you cannot be too careful and 10 means that most people can be trusted.”	11-point scale	36.5	38.4	34.8	38.3	35.3

Note: For the questions using a dichotomous response scale, numbers indicate the proportion of the population indicating trust. For the questions using a 11-point response scale, numbers indicate the proportion of the population indicating a response between 7 and 10. * denotes significance at the 10% level.

Source: ONS (2016), “Statistics on trust for methodological testing from the opinion’s survey, Oct 2015 to May 2016”, Office for National Statistics, UK, www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/adhocs/006326statisticsontrustformethodologicaltestingfromtheopinionsurveyoct2015tomay2016.

StatLink  <http://dx.doi.org/10.1787/888933584165>

When asked about trust with the caution rider version, fewer women reported that most people can be trusted than was the case for men (30.7% vs. 33.6%). By contrast, when using the question wording without the caution rider, more women (40.4%) reported a score of 7 to 10 than was the case for men (36.8%). A comparable pattern can be observed for older people (over 45 years) vs. younger people (aged 16-44): although a similar proportion of those aged 16-44 reported high levels of trust under both question versions, the same was not true for the older age group who, like women, reported lower levels of trust when the caution phrasing was present in the question.

In a follow-up experiment (also portrayed in Table 3.2), the phrasing of the 11-point scale item was changed slightly to also include a caution rider: “Generally speaking, would you say that most people can be trusted, or that you can’t be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can’t be too careful and 10 means that most people can be trusted”. This effectively reversed the results of the first experiment with regards to the 11-point scale. Whereas before women and older people were more likely to report higher levels of trust than the total population, with the caution rider now included, women (34.8%) and older people (35.9%) reported lower levels of trust than the total population (36.5%). Older people and women can arguably be considered as what Hooghe and Reeskens termed relatively more “vulnerable” to the actions of other

people.² It therefore seems plausible to conclude that interpersonal trust questions that use a *cannot be too careful* phrasing, compared to more neutral question wording that focuses solely on trust, induce a priming effect on relatively vulnerable groups. Resulting responses might reflect differences in cautiousness rather than trust.

When it comes to institutional trust, almost no studies have addressed issues of different question wording. A notable exception comes from the same ONS experiment in 2015/16 that has been described above. Here, two issues have been tested: one regarding whether specifying the context in which trust occurs makes a difference, and one regarding whether the word *trust* can be used interchangeably with the word *confidence*.

In the first experiment (see Table 3.3), respondents were presented with an **A trusts B** type question “I am going to name a number of organisations. For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all?” vs. an **A trusts B to do X** type question “I am going to name a number of organisations. For each one, could you tell me how much confidence you have in them *to act in the national interest*: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all?”. The questions were applied to

Table 3.3. **Comparison of confidence in institutions versus confidence in institutions to act in the national interest**

Institution	Question wording	
	Confidence (%)	Confidence to act in the national interest (%)
Armed Forces		
	A great deal	54.6
	Quite a lot	37.3
Police		
	A great deal	22.1*
	Quite a lot	50.9
Justice system		
	A great deal	10.7
	Quite a lot	44.8
Parliament		
	A great deal	3.8
	Quite a lot	25.4
Civil service		
	A great deal	9.2
	Quite a lot	51.6
National health service		
	A great deal	36.7
	Quite a lot	46.8
Banks		
	A great deal	8.5
	Quite a lot	39.8
	Combined	48.3**
Media		
	A great deal	2.3
	Quite a lot	20

Note: Responses recorded on a 4-point scale: “a great deal”, “quite a lot”, “not very much” or “not at all”.

* denotes significance at the 10% level.

** denotes significance at the 1% level.

Source: ONS (2016), “Statistics on trust for methodological testing from the opinion’s survey, Oct 2015 to May 2016”, Office for National Statistics, UK, www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/adhoc/006326statisticsontrustformethodologicaltestingfromtheopinionssurveyoct2015tomay2016.

StatLink  <http://dx.doi.org/10.1787/888933584184>

a range of different institutions, namely the armed forces, the police, the justice system, Parliament, the civil service, the National Health Service, banks and the media. The results, displaying different response distributions between the two question versions depending on the type of institution, suggest that adding a specific trust context can lead to a slightly different interpretation of the trust measure: adding to *act in the national interest* results in a higher proportion of respondents indicating a *great deal of confidence*, the highest possible category, for all tested institutions except the banks, compared to the unspecified question version. In stark contrast, only 5.5% of respondents indicated a *great deal of confidence* in the banks to *act in the national interest* version (compared to 8.5% in the unspecified question). This trend was even more obvious when considering the proportion of respondents indicating *quite a lot of confidence* in the banks: while 39.8% selected this choice for the unspecified question version, only 26.7% did so with the *to act in the national interest* item. It is also worth noting that the banks are the only institution for which both of these categories declined.

It could sensibly be argued that citizens may very well be relatively confident in banks to manage their money, but very much less likely to believe that financial institutions, in view of the financial crisis that affected the UK sample, *act in the long-term interest of the country*. This demonstrates that, at least for banks in a context of financial crisis, there is indeed a conceptual gap between *confidence* vs. *confidence to act in the national interest* for respondents and that this can make a difference to response distributions. It is not clear yet whether *to act in the national interest* is the best specification (or indeed better than no specification at all), and the impact of possible alternatives (e.g. *to act in my personal interest* or *to do what is right*) on the different types of institutions should be empirically examined in the future. It will also be important to test such specifications in other cultural contexts – the current results might apply to the UK only.³

In a second experiment, the ONS explored the distinction between the concepts of *confidence* vs. *trust* in institutions, with half the respondents being asked: “I am going to name a number of organisations. For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all?” and the other half being presented with the question: “I am going to name a number of organisations. For each one, could you tell me how much you trust that institution: is it a great deal of trust, quite a lot of trust, not very much trust or none at all?” Although the theoretical literature on trust in institutions has suggested that confidence and trust tap into slightly different concepts (see Roberts and Hough, 2005 who claim that trust is something one does and is more concrete, whereas confidence is something one has and is more abstract), the experiment found no clear-cut evidence that this distinction is mirrored in how respondents actually respond to questions. As Table 3.4 shows, no consistent logical pattern between the two question versions appears – while it might be somewhat comprehensible that trust is lower than confidence in the media, it does not seem intuitively apparent why the justice system is the only institution for which respondents report higher trust than confidence. Moreover, if trust and confidence really capture two distinct concepts, it seems quite implausible that the banks, which have exhibited quite polarising results in the *confidence* vs. *confidence to act in the national interest* testing described earlier, have a similar proportion of respondents indicating *trust* and *confidence*.

It thus seems that, when distinctions between two concepts are too narrow, respondents cannot draw a meaningful separation between them. The issue of confidence vs. trust touches upon cultural comparability as well. For internationally valid measures, it is important to keep descriptors broad enough to be easily translatable (OECD, 2013). In

Table 3.4. **Comparison of confidence in institutions versus trust in institutions**

Institution	Question wording	
	Confidence (%)	Trust (%)
Armed Forces		
	A great deal	58.5
	Quite a lot	36.6
	Total	95.2
Police		
	A great deal	34.3
	Quite a lot	46.4
	Total	80.7
Justice system		
	A great deal	12.8
	Quite a lot	45
	Total	57.8
Parliament		
	A great deal	4.1
	Quite a lot	24.8
	Total	28.9
Civil service		
	A great deal	11.2
	Quite a lot	54
	Total	65.2
National health service		
	A great deal	41.3
	Quite a lot	45.3
	Total	86.7
Banks		
	A great deal	9.1
	Quite a lot	44.9
	Total	54
Media		
	A great deal	5
	Quite a lot	23.8
	Total	28.8

Note: Responses recorded on a 4-point scale: “a great deal”, “quite a lot”, “not very much” or “not at all”.

Source: ONS (2016), “Statistics on trust for methodological testing from the opinion’s survey, Oct 2015 to May 2016”, Office for National Statistics, UK, www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/adhoc/006326statisticsontrustformethodologicaltestingfromtheopinionsurveyoct2015tomay2016.

StatLink  <http://dx.doi.org/10.1787/888933584203>

contrast to Anglophone countries, most other cultures do not even distinguish between the two terms of trust and confidence. For instance, the French, Spanish and German languages have only one word for trust (*confiance*, *confianza* and *Vertrauen*, respectively). A similar point regarding the distinctions between different types of trust (in different types of people and different institutions) was made in Chapter 2: while a few distinct sub-dimensions within each of the two main categories of interpersonal and institutional trust are identified by respondents, many of the even finer distinctions made between different types of trust are not very empirically informative.

Key messages

- There is convincing evidence that respondents understand questions about trust measures well, mainly based on their strong validity with regard to real world outcomes. This is especially true for interpersonal trust.

- The exact question wording matters for trust measures. Questions should choose wording that does not refer to concepts other than trust and that is specific and precise to the situation of interest. Regardless of the eventual approach to the question wording adopted, standardisation is key to ensure comparison over time and between groups/countries.
- For interpersonal trust, the standard Rosenberg question's introduction of the caution concept impacts on the distribution of responses, with more vulnerable groups (e.g. women and the elderly) reporting lower trust with this phrasing. Hence, a more neutral question wording is preferable for interpersonal trust.
- For institutional trust, specifying the context of institutional behaviour can make a difference in some cases. For example, in the case of financial institutions in contexts which were affected by the financial crisis, adding the specification to *act in the national interest* significantly impacts on respondents' evaluations. When planning other trust questions, it is worth further investigating which other specifications (e.g. *to act in my personal interest* or *to do what is right*) matter for which institutions, and whether a specification should be used at all.
- If concepts are too narrowly related, respondents might have difficulty differentiating between them, as is shown by the analysis of the phrasing of *trust* vs. *confidence* in institutions in the ONS sample. Question wording should be precise enough to be understood by respondents, without getting into extremely subtle nuances (which might also pose an issue for translatability across countries).

3.4. Response formats

The response format chosen for attitudinal measures such as trust questions can have a non-trivial effect on the validity, reliability and comparability of responses. The format needs not only to properly represent the construct of interest and represent the full range of possible responses (including no-response options) but also to be understandable to respondents in the sense that they can provide meaningful and consistent replies. Aspects of the response format considered in the following section include the scale length, scale labelling and response order. Where possible, evidence specific to trust questions will be drawn upon, otherwise lessons from the broader literature on attitudinal questions are emphasised.

The issues

Scale design deals with the maximisation of discriminatory power or sensitivity based on choosing an optimal range of response options for the concept at hand. A good scale will capture as much meaningful variation between responses as exists (which is especially important for single-item measures). Offering too few response options might lead to an inability to detect variation and frustrate respondents, who might feel their attitudes cannot be accurately expressed. However, respondents might be equally burdened by too many response categories, especially if the categories are too close for them to distinguish between cognitively.

Scale labelling, or the way in which response options are described, can influence responses by setting a reference frame about the expected range within which responses can fall. The choice of scale labels, including the anchors at the scale ends, therefore needs to reflect the full range of possible response categories, without compromising reliability by offering meaningless response options, which can lead respondents to respond randomly. Lastly, especially in combination with the selected survey mode, the order of the response options offered can lead to certain options being selected by default by satisficing respondents.

The evidence

When it comes to determining the type of scale and the optimal number of response options being offered to respondents, a couple of general rules have been established in the broader literature on self-reported measurements. For example, if all response categories have verbal labels, five categories are thought to be the maximum number that respondents can process without visual cues (Bradburn et al., 2004). If the response categories are numerical and anchored by verbal descriptions at the two scale ends, respondents tend to respond more successfully to longer scales, as only the end-points need to be memorised. Here, psychometric evidence suggests that the majority of respondents can reliably distinguish between more than six or seven categories (*ibid.*). In general, longer numerical scales (with verbally-labelled anchors at the ends) have been found to enhance test-retest reliability and internal consistency among attitude measures (Weng, 2004; Alwin and Krosnick, 1991; Scherpenzeel and Saris, 1997; Cummins and Gullone, 2000), and validity is likely to improve with an increasing number of numerical scale points (Preston and Colman, 2000). By the same token, Preston and Colman (2000) have also found the internal consistency and discriminatory power of scales with few items to be low in general. Another body of research has posed the question whether to use an odd or even number of response categories, the former including a natural scale mid-point. Chang (1994) argues that – while a mid-point can result in unmotivated or undecided respondents clustering their replies around the middle category – the absence of a mid-point forces respondents to declare an attitudinal preference, even where one might not exist, hence resulting in random responses. In the case of bipolar attitudes (*yes/no* or *agree/disagree*), several studies support an odd number of response options to provide a mid-point that gives respondents the ability to express a neutral position (Alwin and Krosnick, 1991; Bradburn et al., 2004).

The available evidence on aspects of scale labelling suggests that scale anchors can encourage specific response biases. For instance, the use of *agree/disagree* or *yes/no* has been linked to acquiescence bias or “yea-saying” regardless of the actual question content, or even to encouraging socially desirable responding (Krosnick, 1999). Further, it appears to be sensible to adopt absolute scale anchors (e.g. *completely/not at all*) to clearly mark scale ends and offer the full spectrum of possible attitudes to respondents (OECD, 2013). In terms of labelling the points along a scale, there is some debate about whether this should be done numerically or verbally. On the one hand, a few studies have argued that adding verbal labels to all numbered response categories can produce more reliable and stable responding by clarifying the meaning of the different categories to respondents (see Alwin and Krosnick, 1991; Pudney, 2010). On the other hand, contradicting evidence suggests that numerical scales are more accurate than verbally labelled ones and can help to convey scale regularity and equally spaced intervals to respondents (Newstead and Arnold, 1989; Maggino, 2009). Furthermore, numerical labelling allows for the use of a greater number of response options, which is of particular importance for single-item measures. Lastly, in the interest of international comparability, numerical scales are much less likely to pose translation challenges across different languages and contexts (OECD, 2013). Overall, as scale labelling can have a non-negligible impact on the distribution of responses, it will be essential to be consistent in the approach to labelling once one is adopted.

Not only can the number of response options and their labelling characteristics influence answers, but so can the order in which they are presented. Response order effects can be categorised as primacy effects (where satisficing respondents are more likely to select the earlier response options in lists) and recency effects (where respondents tend to choose

the later response options). While primacy effects tend to occur more when response options are presented visually (which can be a useful way to guide respondents), recency effects are more common when choices are read aloud by interviewers (Krosnick, 1999). Both types of order effect become stronger due to item difficulty and respondent fatigue, and respondents with low cognitive abilities are more prone to exhibit this type of response bias (*ibid.*). In general, these effects are more likely to pose a problem for scales that are fully verbally-labelled rather than numerical. However, even numerical response order should be presented consistently (i.e. 0-10 instead of 10-0) in order to minimise mental switching between positive and negative normative outcomes (OECD, 2013).

How does some of this generally applicable advice translate to trust measures? Until recently, the most commonly used measure of interpersonal trust – the Rosenberg question – has been dichotomous. This measure has been used in a large number of surveys, most notably the Civic Culture surveys, the American National Election Studies from 1964 to 1976 and from 1992 to the present, the General Social Survey from 1972 to the present, and the World Values Surveys (WVS) since 1980. In the late 1990s, the Swiss Household Panel and the Citizenship, Involvement and Democracy (CID) cross-national surveys in Europe shifted to a 0-10 numerical point scale. Shortly thereafter the European Social Survey (ESS) adopted the 11-point scale as well.⁴

If one considers the scale length that respondents themselves seem to prefer, Lundasen (2010) conducted qualitative pretesting of World Values Survey interpersonal trust questions on a small Swedish sample where he encouraged the respondents to think aloud while considering the items. The respondents expressed difficulty in grading items on scales with four different points or more and tended to prefer questions with dichotomous answers. However, since the sample in this study was extremely small, it is doubtful how much weight should be given to its findings, especially in light of the larger quantitative research available on the performance of the two types of scales.

On the one hand, Uslaner, while acknowledging that dichotomous trust measures are not without problems, states that the dichotomous question is well understood when people are asked what it means, and it is stable both over time when asked in panels and from parents to children (Uslaner, 2002, pp. 68-74). He also holds that while the arguments for the greater precision of the 11-point numerical scale used in the ESS and CID for both institutional and interpersonal trust seem compelling, respondents are confused by and unable to process the larger number of response options (Uslaner, 2009). Using the ESS and CID surveys for the US, Romania, Moldova and Spain (the last three of which include both dichotomous and 11-point-scale trust measures in institutions and other people), he finds what he calls “evidence of a flight to the middle” or “systematic clumping”. More than 40 per cent of all answers to the 11-point scale questions across the multiple surveys he considers concentrate around the middle of the distribution, namely around the 4 to 6 values. However, it is very debatable whether this response pattern can actually be interpreted as problematic. There is no empirical evidence that middle values do not actually reflect how respondents feel. On the contrary, items that do not include mid-points, such as the dichotomous measure, might actually force respondents to express an attitude that does not reflect their true preference.

On the other hand, at least three studies have explicitly spoken in favour of the 11-point scale over the dichotomous scale for trust measures. Hooghe and Reeskens (2008) analysed data from the 2006 Belgian Youth Survey, which included both measures of

interpersonal trust (with about 40 questions between both items to prevent carry-over effects). Using regression analysis with respondents' demographic data, they found that the 11-point measure of trust, unlike the dichotomous measure, predicts involvement in voluntary associations, which has been linked to trust theoretically in the social capital literature (see Stolle, [1998. However, these findings need to be interpreted in the light of the sample consisting of 16-year-olds and the authors using a combined multi-item measure of trust and fairness (as envisaged by the original Rosenberg scale). Zmerli and Newton (2008) argue that there are theoretical reasons to assume that interpersonal and institutional trust go together, yet many studies find no significant correlations between the two constructs. The authors argue that this is because the majority of this research is based on surveys with short trust scales (4-point scales or dichotomies), such as the World Values Survey and the Eurobarometer. In fact, three of the four cross-sectional studies that find significant correlations between interpersonal and institutional trust use 11-point rating scales (Jagodzinski and Manabe, 2004; Denters, Gabriel and Torcal, 2007; Zmerli, Newton and Montero, 2007). Drawing on 24-country data from the ESS and US CID surveys that include the three Rosenberg interpersonal trust questions as well as six institutional trust questions (all on an 11-point scale), Zmerli and Newton themselves find a strong and robust correlation between interpersonal trust and confidence in institutions after controlling for various socio-economic factors that have commonly been found to be associated with the two types of trust. The results stand even when the three-variable Rosenberg scale is replaced with the single 11-point measure for generalised trust. Saris and Gallhofer (2006) come to similar conclusions when examining the British pilot study of the second wave of the ESS survey, which asked the same questions about interpersonal trust and institutional trust twice, once with a 4-point rating scale and once with an 11-point scale. In both cases, the three questions of the Rosenberg scale were used to measure interpersonal trust, and three questions about parliament, the legal system and the police were asked to tap confidence in public institutions. Only one of the correlations between interpersonal trust and institutional confidence was statistically significant in the case of the 4-point scale, but all nine were significant for the 11-point scale. Neither study, by Saris and Gallhofer or by Zmerli and Newton, addresses whether the correlations found could be due to shared method variance, and further research in this area will be needed to draw definite conclusions.

Key messages

- Different response options lead to different and not necessarily interchangeable measures. Therefore, a standardised approach to response format to ensure the consistency of measurement, especially in an international context, is highly advised.
- The available evidence in terms of general studies and specific information from trust measures suggests that a numerical 11-point scale with verbal scale anchors is preferable over the alternatives, as it allows for a greater degree of variance in responses and increases overall data quality as well as translatability across languages.
- Numerical response order should be presented consistently (i.e. 0-10 instead of 10-0) in order to minimise mental switching between positive and negative normative outcomes.
- When choosing scale anchors, the labels should represent absolute responses (e.g. completely/not at all) to minimise acquiescence bias and socially desirable responding and to allow for the full spectrum of possible responses.

3.5. Survey context

Results can be influenced not only by the design of the individual item itself, but also by the wider survey context in which the item is situated. Since individual survey questions are not asked in isolation, but as part of a continuous flow of items, the position within a sequence of items may conceivably influence responses. Apart from the question order within both a module and the larger survey, the broader temporal context in which the survey takes place is another defining feature of the survey context. While the survey mode, or the way in which data is collected, can also be categorised under survey context, this aspect will be addressed separately in the next section.

The issues

Since due to the very nature of language, words and sentences take part of the meaning from the environment in which they occur (Searle, 1979), the survey context can potentially influence how respondents understand and contextualise individual questions, as well as which information is brought to their minds when forming the answers to items. This can include the way questions or modules are introduced, as well as the nature of the questions asked immediately before trust measures. Attitudinal measures have been linked to respondents being likely to construct answers on the spot (as opposed to systematically retrieving specific memories), making them especially prone to context effects (Sudman, Bradburn and Schwarz, 1996; Schwarz and Strack, 2003).

For example, if questions about (potentially negative) experiences with other people are asked just before interpersonal trust questions, or questions about a specific type of government service are posed before general institutional trust items, a respondent might misinterpret what is intended by the actual trust question. Either consciously or sub-consciously, he or she might answer a question mainly in relation to the subject of the preceding items rather than based on the intended focus for the question, an effect often referred to as priming. Priming can influence both the mean level of a measure and the distribution of data if it impacts subgroups within a population differently, hence impairing comparability across surveys and between groups within the same survey. Further, for trend studies that are interested mainly in marginal changes, order effects are important even if they were to shift results only slightly. Depending on the nature of the priming, order effects can also suppress or inflate correlations between the variables, therefore putting substantive conclusions about their relationships in doubt.

The placement of trust questions within the larger survey context thus can strongly influence which information respondents take into account when constructing their answers, and order effects should be kept in mind when designing questionnaires. Apart from priming through the survey content itself, what is in the top of their minds when respondents formulate answers on their trust level could also be affected by the larger context in which a survey takes place, such as amidst major political scandals or events that could infer perceptions of interpersonal trust, such as a terrorist attack or the exposure of fraud schemes.

The evidence

The literature on the effect of question order on responses distinguishes between two possible directions of influence that can occur, namely assimilation and contrast effects. Assimilation effects are thought to occur when responses are consistent with the information being made salient through a priming question, while contrast effects take place when a response contrasts with the previous information (OECD, 2013).

Schuman and Presser (1996) point out that while order effects for attitudinal questions are not a rare occurrence, it is not the case that *any* disruption of sequence changes responses, and due to the possibility of either assimilation or contrast effects occurring, order effects can actually be hard to predict. Merely placing two items with similar content next to each other does not necessarily create an order effect. Only if respondents have a need to make their answer to the second question consistent with their answers to the first will such an order effect be created. Therefore, more targeted methodological research will be needed to discover what type of question or what type of context triggers which effect. However, a couple of general suggestions about how to deal with question order effects can be derived from experiments across 34 different US surveys (e.g. the Detroit Area Study and the NORC General Social Survey) in the 1970s. Firstly, context effects occur most often when two or more questions deal with the same or closely related issues (*ibid.*). It thus seems logical that they could be prevented if such items are separated within the questionnaire and are buffered by either intervening text or questions. Deaton (2011) found that the influence of political questions on life evaluations was reduced when a “buffer” question between institutional questions and life evaluations was added. More systematic research on the actual impact of transitional text or questions will be needed, as their effect will depend on whether the buffer items are related or unrelated to the actual questions of interest (Schwarz and Schuman, 1997). Furthermore, since the smooth organisation of a survey usually groups similar items together for reasons of coherence and to reduce the response burden, a compromise between avoiding order effects and maximising the ease of interviewing will have to be struck. The second general rule that Schuman and Presser establish states that general summary-type questions that ask respondents to evaluate complex topics in an overall way seem to be more sensitive to position than are more specific questions. This speaks for moving from a broad to a narrow level of specificity, e.g. by placing items about general interpersonal trust before questions about trust in specific groups, or starting with trust in government before asking about specific institutions.

Only very few studies have dealt with order effects in the specific context of trust measurements. In the case of institutional trust, Smith (1981) used a split-sample experiment to test the impact of a set of political alienation questions on subsequent institutional trust items and found that only the very first item (on confidence in major companies) showed a significant decline in trusting responses. This might be an indication of what is called a salience effect, where the first item in a list is more likely to be affected by order effects. Such salience effects will have to be taken into account whenever a list of items is used, for example in the case of institutional trust modules that ask about various organisations. This finding also highlights the need for consistency of order within question modules across surveys and over time.

In the case of interpersonal trust, Smith (1997), who examines the Rosenberg scale (which includes the general interpersonal trust question alongside items on the fairness of other people and the likelihood of their taking advantage of the respondent) in the US General Social Survey, states that “these items are prone to context effects because they call for global assessments of people in general based presumably on one’s entire life experience. Making judgements based on such massive, cognitive retrievals are difficult and open to variability. Sampling of one’s own memories on such broad topics tend to be biased rather than complete or random. Questionnaire context is one factor that biases the cognitive processing and in turn influences the summary judgments” (p. 174). Specifically, Smith found that trusting responses decline by 7.7% when the question was preceded by

items on negative experiences (crime and victimisation) rather than by questions on political ideology, the equalisation of wealth, divorce laws or the legalisation of marijuana.

Very interestingly, the direction of influence between attitudinal and experience measures can also run the opposite way: the US National Crime Survey, conducted from 1972 to 1975, included a module on the victimisation experience of respondents, which was placed after an attitudinal module about crime for a random half of the sample. For this subsample, victimisation reports increased significantly compared to when the attitude module was omitted (Gibson et al., 1978; Cowan, Murphy and Wiener, 1978). A plausible explanation is that the attitude items stimulated memory for and willingness to report on victimisation experiences (which are usually considered to be factual).

This highlights two potentially important messages for trust measures: first, different types of trust measures, ranging from evaluative to experience-based, can be prone to order effects. Placing them as early as possible in the survey should mitigate interference from other questions. It should be noted that as early as possible does not imply at the very beginning, without giving the interviewer the chance to build some degree of rapport with the respondent. Overall, it will be most important to avoid asking trust questions immediately after items that are likely to elicit strong emotional responses or that refer to experiences with other people or institutions. Second, it will be essential to consider not only the priming effect on the trust items in question, but also the priming effect these questions themselves can have on subsequent items, especially if they deal with similar content.

Priming can also occur with regard to broader environmental effects other than those concerning the immediate survey context. For example, in the case of subjective well-being measures, Deaton (2011) found that impactful short-term events, such as major news events (in his case, the 2008 financial crisis) and seasonal holidays, were associated with specific bumps in US time series data. There could be good reasons to assume that such events also impact mood and responses to interpersonal and institutional trust questions. Although financial crises and events such as terrorist attacks typically occur unexpectedly, regular events such as holidays, religious festivities and elections can be taken into account in survey scheduling and should probably be avoided. As with subjective well-being measures, it would be preferable to stage data collection throughout the year or at least over multiple days and weeks to minimise the influence of external events on responses (OECD, 2013).

Key messages

- Although order effects do not appear in every case and every survey, they can have a significant impact on responses when they do and should not be dismissed lightly.
- Order effects occur most often when two or more questions deal with the same or closely related issues, and initial evidence backs a mitigation strategy that either separates trust items within the survey as much as possible without destroying the coherence of the questionnaire or uses intervening text as a buffer.
- Whenever lists of trust items are used, two rules apply: first, general summary-type questions that ask respondents to evaluate complex topics in an overall way seem to be more sensitive to position than are more specific questions. Therefore, a survey should move from a broad to a narrow level of specificity within a group of questions, e.g. by placing items about generalised trust before questions about limited trust. Second, in order to control for salience effects (where the first item in a list is more likely to be

affected by order effects), which is especially important when using a list of different institutions, the order of items should be randomised.

- Trust measures should be placed early enough in the survey to avoid interference from other questions, but late enough to allow for bonding between interviewer and respondent. Overall, it will be key to avoid asking trust questions immediately after items that are likely to elicit strong emotional responses or that refer to experiences with other people or institutions. Questionnaire designers should also not forget about the potential effect that trust questions themselves can have on subsequent items, in particular those dealing with similar content.
- In order to minimise the impact of the broader external survey context, including holidays, seasons and elections, it would be preferable to stage data collection throughout the year or at least over multiple days and weeks.

3.6. Survey mode

Surveys can be conducted in a variety of ways. These include self-administered questionnaires (SAQs), traditionally conducted in a pen-and-paper format, but which increasingly involve internet-online surveys; computer-assisted self-interviews (CASI); telephone interviews and computer-assisted telephone interviews (CATI); and pen-and-paper interviewing (PAPI) and computer-assisted personal interviews (CAPI), usually conducted through visits to the survey respondent's home (OECD, 2013). Recent years have also seen a rise in mixed-method data collection, combining several of the above-listed modes in the same survey. The main distinction between the different survey modes is usually drawn between self-administered or interviewer-led surveys (Holbrook et al., 2003; Metzger et al., 2000; Turner et al., 2005; Tourangeau and Yan, 2007). Different survey modes can substantially influence how respondents process and reply to questions, as well as how much information they feel comfortable to reveal.

The issues

In practice, the choice of survey mode will be influenced by a variety of factors, including coverage and availability of sample frames, financial costs and fieldwork time, as well as the suitability of the questionnaire (Roberts, 2007). On top of that, the potential for error caused by survey mode should also feature in the survey mode selection process for trust questions. Coverage error, non-response error, measurement error and processing error can all be influenced by survey mode (*ibid.*).

This section focuses on the potential for measurement error, namely satisficing, or the use of response biases and heuristics by respondents, and sensitivity. All survey modes vary substantially in the pace with which the survey is conducted, the extent to which the flow of questions is determined by the interviewer or the respondent, whether a respondent can revisit questions throughout the survey, which types of visual aids are presented to the respondent, the extent of human interaction between interviewer and respondent, as well as the privacy of answers. These dimensions can substantially impact how respondents understand questions and portray themselves.

There are theoretical reasons to assume that there is a higher risk of satisficing in self-administered vs. interviewer-led modes (see Roberts, 2007). In self-administered modes there is no time pressure but also no interviewer to facilitate or motivate. Especially in

online surveys, multitasking is possible. Therefore, the overall cognitive burden and risk of satisficing in self-administered surveys is higher.

Since the concept of sensitivity has not been introduced yet in this chapter, a few explanatory words are warranted. The mode, especially through the degree of privacy that it affords, might cause certain types of self-presentational behaviour by respondents, sometimes with and sometimes without their own knowledge. For example, respondents might want to appear consistent across answers (e.g. if a respondent answers first that he or she approves of the current political leadership, he or she might also be tempted to indicate being satisfied with democracy and trusting in public institutions later on so as not to appear erratic). Social desirability is another example of self-presentational behaviour and relates to the difficulty of truthfully reporting an attitude or behaviour that violates existing social norms and may be deemed inappropriate by society. To conform to social norms, respondents may present themselves in a positive light, independently of their actual attitudes and true behaviours. More specifically, respondents might tend to admit to socially desirable traits and behaviours and to deny socially undesirable ones. This is an issue for data quality, as in the case of socially undesirable activities, sample proportions will underestimate the true prevalence and frequency of the attitude while simultaneously overestimating the true level of socially desirable behaviour. This might be an issue, specifically for questions on trust, if it is societally frowned upon to openly declare distrust of specific population groups, or if it is the prevailing fashionable norm to be sceptical of public institutions.

Social desirability is a distinct aspect of what can be termed the “sensitivity” of questions. While the issue associated with social desirability is the sensitivity of an answer, the sensitivity of the question itself poses a challenge if the question topic is deemed to be intrusive (too private or taboo in the culture of the respondent) or if it would be risky to the respondent if his or her true answers were made public or known to third persons beyond the survey setting. The costs and negative consequences could include prosecution or job loss. The reason for their sensitivity is likely to be different for interpersonal and institutional trust: it is plausible to assume that while items involving interpersonal trust are more likely to be affected by social desirability bias (it might not be considered socially acceptable to not trust other people, in particular members of another religion or nationality), while threat of disclosure might be an issue especially for questions about trust in institutions, particularly if the item is included in an official survey conducted by the government or its statistical agency. In terms of the effect on data quality, questions that are too intrusive or pose a threat if disclosed can increase the unwillingness to reply or result in a large number of missing values and *don’t know* responses or in an overestimation of true attitudes if respondents fall back on the “publicly safe” answer.

A common approach to assess whether certain questions are prone to sensitivity-related response biases is a post-hoc assessment via empirical indicators of survey quality such as item non-response rates (Lensvelt-Mulders, 2008). Chapter 2 of these Guidelines has already analysed the item-specific non-response rates for trust questions for the Gallup World Poll and the European Social Survey and highlighted that, while all trust questions perform better than income questions, the bulk of trust questions have higher item-specific non-response rates than more straightforward questions on marital status, education or gender. Institutional trust questions perform worse than interpersonal trust questions in this analysis. Even religion, which is often considered a sensitive topic to ask about, had a non-response rate of less than half of most institutional trust questions. This suggests that trust questions, in particular those concerning institutional trust, can indeed be considered sensitive.

The evidence

While there are very few experimental studies specifically looking at trust and measurement error in relation to survey mode, many lessons from the literature on other self-reported items are applicable here.

Considering the impact of survey mode on the use of response biases and heuristics, there is indeed some evidence supporting the theory that satisficing in its various forms is more likely in self-administered modes than in interviewer-administered modes. For example, higher levels of acquiescence were found in a mixed-mode design that included self-administration than in one which mixed interviewer-administered modes only (Beukenhorst and Wetzels, 2009), and several scholars reported higher levels of *don't know* responses by self-administered (internet) surveys than in interviewer-led surveys (telephone, face-to-face) (see Duffy et al., 2005; Heerwegh, 2009). However, the evidence for satisficing in self-administered modes is not overwhelming – many times, the reported differences were either not significant or significant at very low levels.

Regarding the issue of sensitivity, the tables turn and self-administered survey modes perform much better compared to interviewer-led techniques. Various experimental field studies have established strong evidence that self-administered survey modes, in comparison to interviewer-led techniques, increase the levels of reporting socially stigmatised medical conditions such as depression or sexually transmitted diseases (Villarroel et al., 2008; Krumpal, 2013), socially undesirable activities such as illicit drug use, risky sexual behaviour and abortions (Gribble et al., 1999; Tourangeau and Yan, 2007), as well as socially unaccepted attitudes about race and same-gender sex (Krysan, 1998; Villarroel et al., 2006). Trust-specific evidence pointing in the same direction comes from Cycle 27 of the Canadian General Social Survey on Social Identity, which included questions on interpersonal and institutional trust. Prior to Cycle 27, data collection was done using only computer assisted telephone interviews (CATI); Cycle 27 was the first cycle where an internet self-response option was offered to respondents, and close to 7 000 out of roughly 20 000 respondents completed the survey using the internet option. Using propensity score matching to control for other mode effects (e.g. non-response bias, selection bias and process bias), Statistics Canada compared the responses of both types of data collection and found that CATI respondents showed significantly higher trust scores than did internet respondents.

Thus, while self-administered surveys, in particular web-based ones, carry a slightly higher risk of satisficing, there is a strong case to place trust items in self-administered modules whenever possible, for example in increasingly common mixed-method data collection procedures of national statistical offices (NSOs).

If a face-to-face survey is the only option for data collection, a couple of rules can mitigate the impact of sensitivity-related biases. First, there are documented effects of interviewers' characteristics (e.g. gender and socio-economic status) and assumed interviewer expectations on social desirability bias: Katz (1942) found increased reporting of pro-labour attitudes when interviews were conducted by working-class interviewers. Enumerators should therefore reveal as little of their own social identity as possible during interviews.

Second, several innovative interview methods can be applied to enhance respondents' feeling of privacy: for example, the "sealed envelope technique" (De Leeuw, 2001; Bradburn and Sudman, 1979) involves handing a separate self-administered questionnaire to the respondent in the sensitive questions part of the interview. Respondents are then asked to

complete the questionnaire, place it in an envelope, seal it and return it to the interviewer. Another method, the “unmatched count technique”, involves randomly splitting the sample into two (Biemer and Brown, 2005). One group of respondents is asked to answer a short list of questions that includes only a set of non-sensitive items. The other subsample has to respond to a longer list consisting of the same non-sensitive items plus sensitive ones. Without telling the interviewer which specific items were answered yes, respondents in both groups count the number of positive answers and report solely the sum of these items. An unbiased estimate of the population’s proportion not trusting specific groups or institutions can be obtained by calculating the difference between the two subsample means. While the unmatched count technique is quite innovative and has been successfully applied across a range of stigmatised behaviours (e.g. Haushofer and Shapiro, 2016), it can deal only with yes/no binary response formats at the moment and does not allow for individual-level analysis.

A third way to heighten respondents’ sense of privacy and willingness to co-operate, both in face-to-face and self-administered survey situations, is to highlight confidentiality and data protection assurances at the beginning of the survey. Singer et al. (1995) reviewed the experimental literature on the effects of confidentiality assurances in questionnaire introductions. Although the average effect size was small, the authors found that such confidentiality assurances resulted in lower item non-response and higher response accuracy for sensitive items (including income). While many NSOs might already make use of data protection assurances, it is worth emphasising these even more when trust and other sensitive questions are included in the questionnaire.

Key messages

- While there is some evidence that self-administered surveys carry a higher risk of satisficing, this evidence is neither consistent nor overwhelming.
- There are ways of appropriately tailoring the survey design to reduce social desirability bias and concerns about the threat of disclosure. Sensitivity-related response biases can be reduced by increasing the anonymity of the question-and-answer process (e.g. through self-administered interviews), by decreasing the respondent’s concerns about data protection (e.g. via confidentiality assurances), or by controlling the survey situation (e.g. not having enumerators give out information about their own social identity).
- While placing trust questions in self-administered surveys is strongly preferred, the use of innovative interviewing methods such as the sealed envelope or unmatched count technique could be explored in face-to-face surveys.

3.7. Response styles and the cultural context

The previous sections of this chapter have reviewed various methodological features that can affect response biases and data quality. In addition, individual respondents themselves can consistently be prone to certain forms of response biases or repeatedly rely on particular response heuristics. This type of constant response pattern is known as response style.

The issues

If a respondent consistently relies on a specific style of answering questions, a systematic bias across self-reported variables can be generated. This noise can translate into artificially higher correlations between these self-reported measures, an issue referred

to as common method variance (OECD, 2013). Since all trust measures are self-reported, they are potentially affected by response styles and common method variance.

Respondents are said to be particularly likely to rely on response styles as default patterns of answering if they are fatigued or confused by the way a question is presented or because of lack of knowledge or memory failure. It has been suggested that certain groups of respondents may be more likely to rely on response styles than others, for example people with lower cognitive skills (Krosnick, 1999). Another sometimes-cited factor behind response styles is a respondent's temperament and character (Spector et al., 2000), which can tip a response pattern towards more negative or more positive answers, depending on whether a person is more optimistic or pessimistic.

Beyond individual variation in response styles, a particular concern for international comparisons is the extent to which respondents from different cultures or linguistic groups might exhibit different response styles when answering trust and other self-reported questions. If it can be demonstrated that different countries systematically rely on different response style patterns, including scale use, the accuracy of comparisons between them may be limited.

The evidence

The evidence on response styles is not specific to trust items, but findings from other self-reported measures, including subjective well-being, are applicable and will be drawn upon in the following.

The literature on response styles and subjective measures paints a mixed picture, with some reviews finding a significant influence of acquiescence and other systematic errors on affect measures (see Watson and Clark, 1997). However, a number of other studies suggest that response styles have only a negligible effect on the level of subjective well-being measures (see Moum, 1988; Schimmack, Böckenholt and Reisenzein, 2002). It is also debated whether differences in response styles between population subgroups impact the overall validity of results in practice. For example, some studies have found that “nay-saying” is more common among younger people than older people, as well as among respondents from higher educational backgrounds (e.g. Gove and Geerken, 1977). Importantly, though, these differences did not alter overall relationships between socio-demographic variables (including income, occupation, marital status, race, gender, age and education) and self-reported mental well-being (OECD, 2013).

In general, the biggest challenge to identifying whether a response style is present and to quantifying its impact on results is that response patterns are extremely difficult to verify externally against a common standard or actual behaviour. Usually, response styles are “detected” by comparing a respondent's choices across a variety of survey items: if someone selects the (positive or negative) extremes of scales consistently, or chooses to agree systematically with self-reported statements, he or she is considered to follow an extreme responding or acquiescent response style. For example, Marín, Gamba and Marín (1992) estimate acquiescence through counting the number of times a respondent agreed with a question and then created an extreme responding indicator by counting the times a respondent selected either of the scale anchors. However, unless the responses are logically contradictory (e.g. agreeing to a statement that no one can be trusted and to a statement that everyone can be trusted), it is difficult to tell whether these responses are due to consistent response biases or indeed genuinely reflect the respondent's feelings and level of trust. Thus

we often do not know whether a pattern adds error to the data or represents meaningful variation in trust.

Studies that examine differences in response styles between countries run into similar problems. On the one hand, a couple of works suggest that response styles do indeed vary between cultures. For example, Van Herk, Poortinga and Verhallen (2004) examined marketing data from six EU countries and discovered systematic differences of medium effect size in acquiescence and extreme response styles, with both styles being more prevalent in data from Mediterranean countries (Greece, Italy and Spain) than from northwestern Europe (Germany, France and the United Kingdom). Marín, Gamba and Marín (1992), Clarke (2001) and Holbrook et al. (2006) all suggest that US Hispanics and African-Americans prefer more extreme response categories and are more likely to acquiesce compared to samples of US whites. In contrast, Asian Confucian cultures have been found to be more likely to respond moderately (Lau, Cummins and McPherson, 2005; Lee, et al., 2002) and to be more prone to social desirability bias (Abe and Zane, 1990; Middleton and Jones, 2000) than less collectivist Western nations. However, with the exception of the 2004 Van Herk, Poortinga and Verhallen paper, which includes measures of actual behaviour, all other study designs do not include such validation and therefore do not provide conclusive evidence that the response patterns add noise to the results. Study participants may very well have selected more or less extreme responses and agreed more or less with statements because they represented how they actually feel, rather than how they respond to questions. Furthermore, a recent cross-country study by Exton, Smith and Vandendriessche (2015) drawing on multiple Gallup World Poll waves concluded that culture (including measurement error and actual differences in the experience of life) may account for at most 20% of unexplained country-specific variance in subjective well-being. This effect is small when compared to the role of objective life circumstances in explaining subjective well-being outcomes.

Therefore, both for single and multi-country studies, it is safe to conclude that even in the cases where it can be proven that a response pattern is unambiguously due to differences in response styles and not to actual respondent evaluations, response styles do not seem to harm overall data quality to such a degree that trust and other self-reported measures should be dismissed as invalid. This insight also has implications for the appropriate strategies to deal with response styles: several scholars have suggested either controlling for the factors assumed to drive response styles (e.g. personality) in the analysis or going a step further and applying statistical adjustment techniques such as mean correction or scale standardisation directly to measures believed to be affected by response styles (Greenleaf, 1992; Hofstede, 2001). However, there is a substantial risk that such strategies would eliminate true substantive differences (see Harzing, 2006). Rather than trying to eliminate response bias retrospectively through statistical adjustment, it might therefore be preferable to avoid response style bias in the first place by careful questionnaire design. For example, questions with *agree/disagree* and to a lesser degree *yes/no* response formats might be more likely to prompt acquiescence and should be avoided if possible (Krosnick, 1999). Smith (2003) also suggests that the use of a mixture of positive and negative statements will mitigate both acquiescence and disacquiescence, because it might lead respondents to consider the exact meaning of the question more closely and as a result give more meaningful responses, or at least lead to the responses cancelling each other out. This approach will need to be further tested, as it poses a risk of confusing respondents when the same scale end presents something positive in one item and something negative in a following one. Moreover, questionnaire items containing negations can be difficult to translate into some languages.

In general, given that individuals are assumed to be more likely to rely on response biases and heuristics when they are confused by questions, less motivated, more fatigued and more burdened, the best way to minimise these issues is likely to be through adopting solid survey design principles: avoiding items that are difficult to understand or repetitive or that look too similar; using short and engaging questions that are easy to answer; and keeping respondents interested and motivated (OECD, 2013). These principles are true for all survey measures, and there is no strong reason to assume that trust measures are at any greater risk of eliciting response styles than other self-reported survey items.

Key messages

- Response styles are very difficult to verify externally against a common standard or actual behaviour. More often than not, we do not know whether a pattern adds error to the data or represents meaningful variation in trust.
- Even where the existence of response styles has been established, they do not necessarily seem to harm overall data quality to such a degree that trust and other self-reported measures should be dismissed as invalid.
- Rather than using statistical adjustment techniques to mitigate response style bias, the focus should be on designing the questionnaire so that items are as simple, easy to interpret and minimally burdensome as possible. The overall survey design (including its length and how it is introduced) needs to pay particular attention to respondent burden, motivation and fatigue in order to maximise data quality.
- Question formats that are known to be more prone to response biases should be avoided. A case in point are *agree/disagree* and to a lesser degree *yes/no* response formats, which are more likely to prompt acquiescence.
- For internationally comparative analyses of trust data, one option to get around response style concerns could be to use changes in response patterns over time (including those of different population subgroups) rather than the level of responding.

3.8. Conclusion

This chapter has discussed the various methodological features of question and survey design that can impact measurement error for trust items. Key conclusions and areas warranting future research include the following:

- While trust measures are more sensitive to response biases than more objective measures (such as educational attainment or life expectancy), these biases are also likely to occur in other self-reported measures that are already being collected by NSOs. Although it is essential to be aware of these biases and of the most appropriate question and survey design strategies to mitigate them, the existence of measurement error *per se* is not an argument against gathering data on trust. Especially for items on interpersonal trust, the evidence of their strong validity with regard to real world outcomes demonstrates that these measures are meaningful and worth collecting.
- No matter which approach to question design is adopted by data collectors, standardisation is critical to ensure meaningful comparison over time and between groups and countries.
- The evidence on question wording (especially that drawn from split sample experiments) shows that this is not a trivial matter and that good question wording matters for results. Question wording should avoid referring to concepts other than trust and be specific and precise to the situation of interest. For interpersonal trust, a neutral question wording is

recommended: data collectors should refrain from referring to *caution in dealing with other people*, as this wording can prime more vulnerable groups to report lower trust. For institutional trust, specifying what institutions are expected to do can make a difference in some cases. Overall, question wording should be precise enough to be understood by respondents, without getting into subtle nuances (which might also pose an issue for translatability across countries). If the concepts that different questions try to capture are too narrowly related, respondents might have difficulty differentiating between them (e.g. trust vs. confidence).

- The way answering options are presented can have a significant impact on the distribution of responses. For trust items, a numerical 0-10 scale with verbal scale anchors is recommended, as it allows for a high degree of variance in responses, increases overall data quality and facilitates translatability across languages. The response order should be presented consistently (i.e. 0-10 instead of 10-0) in order to minimise mental switching between positive and negative normative outcomes. The verbal descriptions of the scale anchors should represent absolute responses (e.g. *completely/not at all*) to minimise acquiescence bias and socially desirable responding and to allow for the full spectrum of possible responses.
- Trust measures should be considered within the broader survey context in which they are placed. As with the standardisation of wording and response formats, consistency of order within question modules across surveys and over time is essential to guarantee the quality and comparability of trust measures. Since order effects occur most often when two or more questions deal with the same or closely related issues, trust items should either be separated within the survey as much as possible or buffered by intervening text. Whenever lists of trust items are used, a survey should move from a broad to a narrow level of specificity within a group of questions, e.g. by placing items about generalised trust before questions about limited trust. When thinking about placement of trust questions in the survey, a balance is needed between showing questions early enough to avoid interference from other questions and late enough to allow for bonding between interviewer and respondent. Generally, trust questions should not be asked immediately after items that are likely to elicit strong emotional responses or that refer to experiences with other people or institutions. Questionnaire designers should equally reflect on the potential effect that trust questions themselves can have on subsequent items, in particular those dealing with similar content. Lastly, in order to minimise the impact of holidays, seasons and elections, data collection is recommended to be spread throughout the year or at least over multiple weeks.
- Evidence suggests that trust questions can be sensitive, triggering respondents to answer in a socially desirable way or be unwilling to answer at all. This is especially true for measures of trust in institutions. Self-administered surveys, compared to interviewer-led ones, perform better in terms of minimising social desirability. This benefit outweighs the (relatively weak) counterargument that self-administered surveys carry a higher risk of satisficing. In all survey modes, sensitivity-related response biases can be reduced by decreasing the respondent's concerns about data protection (e.g. via confidentiality assurances) or by controlling the survey situation (e.g. not having enumerators give out information about their own social identity). If face-to-face interviews are the only option, the use of innovative interviewing methods such as the sealed envelope or unmatched count technique could be explored.

- Cross-cultural response styles are very difficult to verify externally against a common standard or actual behaviour. Even where the existence of response styles has been established, they do not necessarily harm overall data quality to such a degree that trust and other self-reported measures should be considered as invalid. If data producers want to mitigate the possibility of response style bias, they should, rather than relying on *ex post* statistical adjustment techniques, focus on designing the questionnaire so that items are as simple, easy to interpret and minimally burdensome as possible. The overall survey design (including its length and how it is introduced) needs to pay particular attention to respondent burden, motivation and fatigue in order to maximise data quality. Moreover, question formats that are more prone to response biases should be avoided: for example, *agree/disagree* and to a lesser degree *yes/no* response formats are more likely to prompt acquiescence.
- Further research is needed on both institutional and interpersonal trust, but especially on the former, for which there is very little methodological evidence available.
 - ❖ First, with regard to question wording for institutional trust, experimental testing should be used to establish which specifications other than *to act in the national interest* (e.g. *to improve the life of someone like me* or *to do what is right*) matter for which institutions. Ideally, these experiments should be carried out across more than just one country.
 - ❖ Second, while it has been suggested that the use of a mixture of positive and negative statements can mitigate both *yay* and *nay* saying, this approach needs to be further tested to rule out the risk of confusing respondents when the same scale end presents something positive in one item and something negative in a following one.
 - ❖ Third, with regard to order effects, it is not yet clear in which cases these occur for trust questions. More targeted methodological research is needed to discover what type of question or what type of context triggers which effect in order to further inform survey design. While there is some evidence that transitional text between questions can act as a buffer to mitigate order effects, various text versions should be tested for their impact on trust questions.
 - ❖ Finally, more research that validates response styles from different cultures against external references, such as actual trusting behaviour in real life or experimental games, would enrich the current body of cross-cultural trust research.

Notes

1. The four interpersonal trust question versions in the ESC survey were: 1) “Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?”; 2) “People can be trusted until they prove otherwise”; 3) “Generally speaking, most people can be trusted. Do you agree or disagree?” and 4) “Generally speaking, you can’t be too careful in dealing with people. Do you agree or disagree?”
2. Evidence of perceived vulnerability of these population groups can be found in responses on whether they feel safe walking alone at night in their neighbourhood, which demonstrate a significantly lower proportion of women and people over 50 years reporting to feel safe compared to other population groups (OECD, 2015).
3. The OECD also partnered with INEGI of Mexico, placing various trust question versions in Mexico’s June 2016 National Survey on Urban Public Security. This study was not a split sample experiment (each respondent was asked two questions within the same survey), and therefore it cannot be ruled out that differences in responses are due to priming or shared method variance. Nevertheless, when the two versions (regular trust in institutions questions vs. *to act in the national interest*) were posed to the 500-person sample, adding to *act in the national interest* did not lead to a strong drop in the

population share which indicated that banks can be trusted *a great deal*. On the other hand, the civil service experienced a drop from 10.6% to 6.8% of respondents indicating that they trust this institution *a great deal*. These results potentially indicate that institutions could carry different connotations in the context of Mexico compared to the UK. Further actual experimental research will be needed to clarify this issue.

4. While some NSOs, such as the Central Statistical Office of Poland, have introduced verbally labelled 4-point scales, comparative methodological evidence on the effect of this type of response scale remains limited.

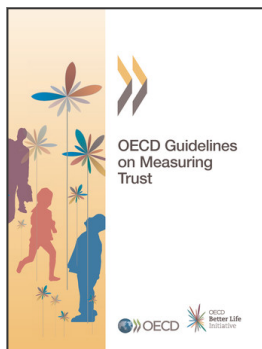
References

- Abe, J.S. and N.W.S. Zane (1990), "Psychological maladjustment among Asian and White American college students: Controlling for confounds", *Journal of Counselling Psychology*, Vol. 37, pp. 437-444.
- Alwin, D.F. and J.A. Krosnick (1991), "The reliability of survey attitude measurement: The influence of question and respondent attributes", *Sociological Methods and Research*, Vol. 20, No. 1, pp. 139-181.
- Beukenhorst, D. and W. Wetzels (2009), "A comparison of two mixed mode designs of the Dutch Safety Monitor: Mode effects, costs, logistics", Paper presented at the European Survey Research Association Conference, Warsaw.
- Biemer, P.P. and G. Brown (2005), "Model-based estimation of drug use prevalence using item count data", *Journal of Official Statistics*, Vol. 21, No. 2, pp. 287-308.
- Bradburn, N.M. and S. Sudman (1979), *Improving Interview Method and Questionnaire Design*, Jossey-Bass, San Francisco.
- Bradburn, N.M., S. Sudman and B. Wansink (2004), *Asking Questions: The Definitive Guide to Questionnaire Design – from Market Research, Political Polls, and Social and Health Questionnaires*, Jossey-Bass, San Francisco.
- Chang, L. (1994), "A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity", *Applied Psychological Measurement*, Vol. 18, pp. 205-215.
- Clarke, J. (2001), "Extreme response style in cross-cultural research", *International Marketing Review*, Vol. 18, pp. 301-324.
- Cowan, C.D., L.R. Murphy and J. Wiener (1978), "Effects of supplemental questions on victimization estimates from the national crime survey", *1978 Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, United States.
- Cummins, R. and E. Gullone (2000), "Why we should not use 5-point Likert scales: The case for subjective quality of life measurement", *Proceedings of the Second International Conference on Quality of Life in Cities*, Singapore.
- Deaton, A.S. (2011), "The financial crisis and the well-being of Americans", *Working Paper*, No. 17128, National Bureau of Economic Research (NBER), Cambridge MA, www.nber.org/papers/w17128.
- De Leeuw, E.D. (2001), "Reducing missing data in surveys: An overview of methods", *Qual. Quant.*, Vol. 35, pp. 147-160.
- Denters, B., O. Gabriel and M. Torcal (2007), "Political confidence in representative democracies: Social capital vs. political explanations", J. van Deth, J.R. Montero and A. Westholm (eds), *Citizenship and Involvement in European Democracies*, Routledge, Abington.
- Duffy, B. et al. (2005), "Comparing data from online and face-to-face surveys", *International Journal of Market Research*, Vol. 47, No. 6.
- Exton, C., C. Smith and D. Vandendriessche (2015), "Comparing happiness across the world: Does culture matter?", *OECD Statistics Working Papers*, No. 2015/04, OECD Publishing, Paris, <http://dx.doi.org/10.1787/5jrappzd9bs2-en>.
- Gibson, C. et al. (1978), "Interaction of survey questions as it relates to interviewer-respondent bias", *1978 Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, United States.
- Gove, W.R. and M.R. Geerken (1977), "Response bias in surveys of mental health: An empirical investigation", *American Journal of Sociology*, Vol. 82, No. 6, pp. 1289-1317.
- Greenleaf, E.A. (1992), "Improving rating scale measures by detecting and correcting bias components in some response styles", *Journal of Marketing Research*, Vol. 29, pp. 176-188.

- Gribble, J.N. et al. (1999), "Interview mode and measurement of sexual behaviors: Methodological issues", *Journal of Sexual Research*, Vol. 36, pp. 16-24.
- Harzing, A.W. (2006), "Response styles in cross-national survey research: A 26-country study", *International Journal of Cross-Cultural Management*, Vol. 6, pp. 243-266.
- Haushofer, J. and J. Shapiro (2016), "The short-term impact of unconditional cash transfers to the poor: Evidence from Kenya", *Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1973-2042.
- Heerwegh, D. (2009), "Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects", *International Journal of Public Opinion Research*, Vol. 21, No. 1, pp. 111-121.
- Hofstede, G.H. (2001), *Culture's Consequences: Comparing Values, Behaviours, Institutions, and Organizations across Nations*, Sage, Thousand Oaks, California.
- Holbrook, A.L., T.P. Johnston and Y.I. Cho (2006), "Extreme response style: Style or substance?", Paper presented at the 61st Annual Meeting of the American Association for Public Opinion Research, Montreal, Canada.
- Holbrook, A.L., M.C. Green and J.A. Krosnick (2003), "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias", *Public Opinion Quarterly*, Vol. 67, pp. 79-125.
- Hooghe, M. and T. Reeskens (2008), "Cross-cultural measurement equivalence of generalized trust: Evidence from the European Social Survey (2002 and 2004)", *Social Indicators Research*, Vol. 85, pp. 15-32.
- Jagodzinski, W. and K. Manabe (2004), "How to measure interpersonal trust? A comparison of two different measures", *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, Vol. 55, pp. 85-98.
- Katz, D. (1942), "Do interviewers bias poll results?", *Public Opinion Quarterly*, Vol. 6, pp. 248-268.
- Krosnick, J.A. (1999), "Survey research", *Annual Review of Psychology*, Vol. 50, pp. 537-567.
- Krosnick, J.A. (1991), "Response strategies for coping with the cognitive demands of attitude measures in surveys", *Applied Cognitive Psychology*, Vol. 5, pp. 213-236.
- Krumpal, I. (2013), "Determinant of social desirability bias in sensitive surveys: A literature review", *Qual Quant*, Vol. 47, pp. 2025-2047.
- Krysan, M. (1998), "Privacy and the expression of white racial attitudes – a comparison across three contexts", *Public Opinion Quarterly*, Vol. 62, pp. 506-544.
- Lau, A.L.D., R.A. Cummins and W. McPherson (2005), "An investigation into the cross-cultural equivalence of the personal wellbeing index", *Social Indicators Research*, Vol. 72, pp. 403-430.
- Lee, J.W. et al. (2002), "Cultural differences in responses to a Likert scale", *Research in Nursing and Health*, Vol. 25, pp. 295-306.
- Lensvelt-Mulders, G.J.L.M. (2008), "Surveying sensitive topics", De Leeuw, E.D., J.J. Hox, D.A. Dillman (eds), *The International Handbook of Survey Methodology*, Erlbaum/Taylor&Francis, New York/London.
- Lundasen, S. (2010), "Methodological problems with surveying trust", manuscript.
- Maggino, F. (2009), "Methodological aspects and technical approaches in measuring subjective well-being", Università degli Studi di Firenze, Working Paper.
- Marín, G., R.J. Gamba and B.V. Marín (1992), "Extreme response style and acquiescence among Hispanics: The role of acculturation and education", *Journal of Cross-Cultural Psychology*, Vol. 23, No. 4, pp. 498-509.
- Metzger, D.S. et al. (2000), "Randomized controlled trial of audio computer-assisted self-interviewing: Utility and acceptability in longitudinal studies. HIVNET Vaccine Preparedness Study Protocol Team", *American Journal of Epidemiology*, Vol. 152, No. 2, pp. 99-106.
- Middleton, K.L. and J.L. Jones (2000), "Socially desirable response sets: The impact of country culture", *Psychology and Marketing*, Vol. 17, No. 2, pp. 149-163.
- Moum, T. (1988), "Yea-saying and mood-of-the-day effects in self-reported quality of life", *Social Indicators Research*, Vol. 20, pp. 117-139.
- Newstead, S.E. and J. Arnold (1989), "The effect of response format on ratings of teaching", *Educational and Psychological Measurement*, Vol. 49, pp. 33-43.
- OECD (2015), *How's Life? 2015: Measuring Well-being*, OECD Publishing, Paris, http://dx.doi.org/10.1787/how_life-2015-en.

- OECD (2013), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264191655-en>.
- ONS (2016), "Statistics on trust for methodological testing from the opinion's survey, Oct 2015 to May 2016", released 10 November 2016, Office for National Statistics, UK, ONS, Newport, www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/adhocs/006326statisticsontrustformethodologicaltestingfromtheopinionssurveyoct2015tomay2016.
- Peadody, D. (1961), "Authoritarianism scales and response bias", *Psychological Bulletin*, Vol. 65, No. 1, pp. 11-23.
- Podsakoff, P.M. et al. (2003), "Common method biases in behavioral research: A critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88, No. 5, pp. 879-903.
- Preston, C.C. and A.M. Colman (2000), "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences", *Acta Psychologica*, Vol. 104, pp. 1-15.
- Pudney, S. (2010), "An experimental analysis of the impact of survey design on measures and models of subjective well-being", *Institute for Social and Economic Research Working Papers*, No. 2010-20, University of Essex.
- Roberts, C. (2007), "Mixing modes of data collection in surveys: A methodological review", *Economic and Social Research Council – National Centre for Research Methods, NCRM Methods Review Papers NCRM/008*.
- Roberts, J. and M. Hough (2005), *Understanding Public Attitudes to Criminal Justice*, Open University Press, Maidenhead.
- Saris, W.E. and I. Gallhofer (2006), "Report on the MTMM experiments in the pilot studies and proposals for Round 1 of the ESS", www.europeansocialsurvey.org/docs/methodology/ESS1_quality_measurement.pdf.
- Scherpenzeel, A. and W.E. Saris (1997), "The validity and reliability of survey questions – A meta-analysis of MTMM studies", *Sociological Methods Research*, Vol. 25, No. 3, pp. 341-383.
- Schimmack, U., U. Böckenholt and R. Reisenzein (2002), "Response styles in affect ratings: Making a mountain out of a molehill", *Journal of Personality Assessment*, Vol. 78, No. 3, pp. 461-483.
- Schuman, H. and S. Presser (1996), *Questions and Answers in Attitude Surveys*, Sage Publications, California, United States.
- Schwarz, N. and F. Strack (2003), "Reports of subjective well-being: Judgemental processes and their methodological implications", D. Kahneman, E. Diener and N. Schwarz (eds.), *Well-being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, New York.
- Schwarz, N. and H. Schuman (1997), "Political knowledge, attribution and inferred interest in politics", *International Journal of Public Opinion Research*, Vol. 9, No. 2, pp. 191-195.
- Searle, J.R. (1979), *Expression and Meaning*, Cambridge University Press, New York, United States.
- Singer, E., D.R. von Thurn and E.R. Miller (1995), "Confidentiality assurances and response: A quantitative review of the experimental literature", *Public Opinion Quarterly*, Vol. 59 No. 1, pp. 66-77.
- Smith, T.W. (2003), "Developing comparable questions in cross-national surveys", J.A. Harkness, F.J. van de Vijver and P.P. Mohler (eds.), *Cross-Cultural Survey Methods*, Wiley, New York.
- Smith, T.W. (1997), "Factors relating to misanthropy in contemporary American society", *Social Science Research*, Vol. 26, pp. 170-196.
- Smith, T.W. (1981), "Can we have confidence in confidence? Revisited", Denis Johnston (eds.), *Measurement of Subjective Phenomena*, US Government Printing Office, Washington, DC.
- Soroka, S., J. Helliwell and R. Johnston (2007), "Measuring and modelling trust", F. Kay and R. Johnston (eds.), *Diversity, Social Capital and the Welfare State*, University of British Columbia Press, Vancouver, BC.
- Spector, P.E. et al. (2000), "Why negative affectivity should not be controlled in job stress research: Don't throw out the baby with the bath water", *Journal of Organizational Behavior*, Vol. 21, No. 1, pp. 79-95.
- Stolle, D. (1998), "Bowling together, bowling alone: The development of generalized trust in voluntary associations", *Political Psychology*, Vol. 19, No. 3, pp. 497-525.
- Sudman, S., N.M. Bradburn and N. Schwarz (1996), *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*, Jossey-Bass, San Francisco.

- Tourangeau, R. and T. Yan (2007), "Sensitive questions in surveys", *Psychological Bulletin*, Vol. 133, pp. 859-883.
- Turner, C.F. (2005), "Reducing bias in telephone survey estimates of the prevalence of drug use: A randomized trial of telephone audio-CASI", *Addiction*, Vol. 100, pp. 1432-1444.
- Uslaner, E.M. (2009), "Is eleven really a lucky number? Measuring trust and the problem of clumping", Manuscript.
- Uslaner, E.M. (2002), *The Moral Foundations of Trust*, Cambridge University Press, Cambridge, United Kingdom.
- Van Herk, H., Y.H. Poortinga and T.M.M. Verhallen (2004), "Response styles in rating scales: Evidence of method bias in data from six EU countries", *Journal of Cross-Cultural Psychology*, Vol. 35, No. 3, pp. 346-360.
- Villarroel, M.A. et al. (2008), "T-ACASI reduces bias in STD measurements: The national STD and behavior measurement experiment", *Sexually Transmittable Diseases*, Vol. 35, pp. 499-506.
- Villarroel, M.A. et al. (2006), "Same-gender sex in the United States: Impact of T-ACASI on prevalence estimates", *Public Opinion Quarterly*, Vol. 70, pp. 166-196.
- Watson, D. and L.A. Clark (1997), "Measurement and mismeasurement of mood: Recurrent and emergent issues", *Journal of Personality Assessment*, Vol. 68, No. 2, pp. 267-296.
- Weng, L.J. (2004), "Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability", *Educational and Psychological Measurement*, Vol. 64, pp. 956-972.
- Zmerli, S. and K. Newton, (2008), "Generalised trust and attitudes toward democracy", *Public Opinion Quarterly*, Vol. 72, pp. 706-724.
- Zmerli, S., K. Newton and J.R. Montero (2007), "Trust in people, confidence in political institutions, and satisfaction with democracy", J. van Deth, J. R. Montero and A. Westholm (eds.), *Citizenship and Involvement in European Democracies*, Routledge, Abington.



From:
OECD Guidelines on Measuring Trust

Access the complete publication at:
<https://doi.org/10.1787/9789264278219-en>

Please cite this chapter as:

OECD (2017), “Methodological considerations”, in *OECD Guidelines on Measuring Trust*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/9789264278219-6-en>

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.