Chapter 2

Methodological considerations in the measurement of subjective well-being

Introduction

The goal of the present chapter is to outline the available evidence on how survey methodology can affect subjective well-being measures and draw together what is currently known about good practice. The chapter focuses on aspects of survey design and methodology and is organised around five main themes: i) question construction; ii) response formats; iii) question context; iv) survey mode effects and wider survey context effects; and v) response styles and the cultural context in which a survey takes place. Each section is structured around the key measurement issues raised, the evidence regarding their impact, and the implications this has for survey methodology.

Much like any other survey-based measure, it is clear that subjective well-being data can be affected by the measurement methods adopted. Maximising data quality by minimising the risk of bias is a priority for survey design. Comparability of data between different survey administrations is another essential consideration. In support of data comparability, this chapter argues in favour of adopting a consistent or standardised measurement approach across survey instruments, study waves, and countries wherever possible, thus limiting the additional variance potentially introduced by differing methodologies.

Perhaps *because* of concerns about their use, quite a lot *is* known about how subjective well-being measures behave under different measurement conditions – in some cases more so than many self-report measures already included in some national surveys. The extensive manner in which subjective well-being questions have been tested offers a firm evidence base for those seeking to better understand their strengths and limitations. However, some questions remain regarding the "optimal" way to measure subjective well-being. National statistical agencies are in a unique position to further improve the evidence base, by answering some of the questions for which large and more nationally-representative samples are required.

The present chapter is framed in terms of the potential for measurement error in subjective well-being data. All measures suffer from error, even the most objective measures used in "hard" sciences. But as argued by Diener, Inglehart and Tay (2012): "We cannot… automatically dismiss findings or fields because of measurement error because science would halt if we did so". Given that some degree of measurement error is inevitable, the goal then is to guide the selection of measures that are *good enough* to enable meaningful patterns to be distinguished from noise in the data. The patterns of greatest interest to policy-makers are likely to include both meaningful changes in subjective well-being over time, and meaningful differences between population subgroups, as well as better understanding the determinants of subjective well-being. These analyses are discussed in greater detail in Chapter 4 (*Output and analysis of subjective well-being measures*) of the guidelines.

In terms of coverage, this chapter considers only the three types of subjective well-being measures set out in Chapter 1: *evaluative* measures regarding life overall; *affective* measures capturing recent experiences of feelings and emotions; and *eudaimonic*

measures. These measures have different properties and, in some cases, show different relationships with determinants (Clark and Senik, 2011; Huppert and So, 2009; Diener et al., 2009; Schimmack, Schupp and Wagner, 2008). By their natures, these measures may also place differing demands on respondents, and may thus vary in their susceptibility to different sources of bias and error.

Wherever possible, this chapter considers general principles of good survey design that will be relevant to all three of these measures of subjective well-being; where issues are of particular concern to one type of measure, this will be highlighted. Where specific evidence on subjective well-being is lacking, the chapter draws on some examples from other literatures (such as those examining attitudes, or subjective perceptions of more objective phenomena such as health). However, it must be emphasised that the scope of these guidelines is firmly focused on measures of subjective *well-being* only, rather than on all possible subjective and/or self-report measures commonly used across a variety of surveys.

The remaining part of this introduction discusses the question-answering process, and the various types of errors or biases that can arise in the course of this process. The sections that follow are then organised according to the various features of survey design that can potentially influence the likelihood of errors or biases. It begins by considering the most narrow design features (question construction and response formats), before broadening out the discussion to question context, placement and ordering. Broader still are mode effects, and other issues arising from the wider survey method, such as the day of the week and the time of year that the survey is conducted. The final section of the chapter then deals with *response styles* (see below) and the cultural context in which a survey takes place – and what that might mean for data comparability, particularly across different countries.

In practice, of course, many of these survey design elements are contingent upon one another – so for example, survey mode influences the question wording and response formats that can be most easily used and understood by respondents. The risks associated with response styles and the wider cultural context in which surveys are conducted can also potentially have implications across most of the other survey design features. Some of these cross-cutting issues and trade-offs are highlighted in Table 2.1. Chapter 3 (an approach to *Measuring subjective well-being*) describes managing the practical trade-offs involved in survey design in more detail, providing advice on issues such as translation and proposing a set of question modules.

The question-answering process and measurement error

In order to answer any survey question, respondents are assumed to go through several cognitive and information-processing steps, which may be performed either sequentially or in parallel. These steps include: understanding the question; recalling information from memory (as relevant); computing or forming a judgement; formatting the judgement to fit the response alternatives; and editing the final answer before delivering it to the surveyor (Schwarz et al., 2008; Sudman, Bradburn and Schwarz, 1996).

Processing and subsequently reporting subjective feelings of well-being may be a novel and potentially demanding task for survey respondents. Evidence from both the time taken to respond to questions (ONS, 2011a), and general non-response rates (e.g. Smith, 2013; ONS, 2011b) suggests, however, that the vast majority of respondents *are* able to provide answers to subjective well-being questions, and usually do so reasonably quickly

Table 2.1. **Possible response biases and heuristics described in the self-report survey literature**

Response bias or heuristic	Expected pattern of responses
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.

(e.g. in around half a minute even for life evaluations). The speed with which responses are provided could be taken to imply that respondents are generally able to complete the task with relatively low levels of difficulty, or it could imply that not all respondents are taking the time to fully consider their responses – and they are relying instead on short-cuts or contextual information to help them formulate their answers. The evidence reviewed in this chapter is largely concerned with the extent to which survey design influences the likelihood of the latter possibility.

Measurement error describes the extent to which survey measures reflect facets other than those intended by the surveyor. This error can either be *systematic*, exerting a bias in the data that is consistent in some way, and that might lead to, for example, values that are systematically higher or lower than might be expected; or *random*, varying from observation to observation 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 it). As well as better understanding how to manage the risk of error through survey methodology, it is also important to understand how survey error might distort results and further analyses (and how this in turn can be managed), which is discussed at greater length in Chapter 4.

Patterns of error – response biases and heuristics

According to Bradburn, Sudman and Wansink (2004), four basic factors can lead to respondent error in self-reported survey measures: i) failures in memory (e.g. material may be forgotten or misremembered); ii) lack of appropriate motivation (e.g. some respondents may be motivated to present themselves in a positive light, or unmotivated respondents may not process questions fully); iii) failures in communication (e.g. the meaning of the

question may be unclear or misunderstood); and iv) lack of knowledge (e.g. respondents may simply not know the answer to the question, but may attempt to answer it anyway). Respondent errors can be either caused or exacerbated by aspects of survey design and context – and different sources of error will often call for different solutions.[1]

Failures in memory, motivation, communication or knowledge are often associated with increased risk of response biases and the use of response heuristics. *Response biases* refer to particular patterns or distortions in how individuals or groups of individuals respond to questions; *response heuristics* refer to (often sub-conscious) rules or short-cuts that respondents may use in order to help them select their answers. Drawing on the classifications of Podsakoff et al. (2003), some examples of response biases and heuristics are described in Table 2.1. Where a respondent exhibits a repeated tendency towards a particular response bias or heuristic, this is referred to as a *response style*.

Patterns of error – contextual cueing

Contextual cueing refers to the influence that subtle cues within the survey context can have on how individuals answer questions (so, for example, a survey issued by a hospital could cue respondents to think about their health when answering questions). The question-answering process is not simply a cognitive, information-processing or memory task: it is also part of a social interaction and a communicative process. Several authors have likened the survey question-and-answering process to a special case of an ordinary conversation (e.g. Bradburn, Sudman and Wansink 2004; Schwartz 1999; Sudman, Bradburn and Schwartz, 1996). This implies that, in the course of establishing the meaning of questions, respondents rely on the same tacit principles that govern other more naturally-occurring forms of conversation – and thus they use, for example, contextual information contained in previous questions to guide the course of communication. Thus, survey design can have unintended consequences for how questions are interpreted.

In order to help them answer questions about subjective well-being, respondents may (consciously or otherwise) seek information about the meaning of a question, and how to answer it, from the wording of the question, the response format, or from other questions in the survey. The survey source and the manner in which questions are introduced can also influence responses, as can the phrasing of the question, which can lead respondents to believe a particular answer is expected (e.g. leading questions that suggest a particular response: "Is it true that you are happier now than you were five years ago?"). Cues also exist in the wider social context: respondents may draw on social norms and what they know about others in order to answer questions about themselves; they may consider their past experiences and future hopes; and they may be reluctant to give answers that they believe to be socially undesirable. There is thus an interaction between social communication and individual thought processes (Sudman, Bradburn and Schwartz, 1996).

To use the language of response biases described in Table 2.1, contextual cues can potentially lead to *priming effects, socially desirable responding, demand characteristics* and use of the *consistency motif*. The wider environment beyond the survey can also influence how respondents feel, and/or what is at the forefront of their minds, even when these are not explicitly discussed during the survey. If factors such as the weather, climate, day of week, and day-to-day events or mood can influence responding, these can also be considered contextual cues. For example, "mood-congruent recall" refers to a phenomenon whereby respondents can find it easier to recall information that is consistent with their current mood (so respondents in a positive mood can find it easier to recall positive information, etc.).

Patterns of error – individual differences

The effects of errors or biases may not be uniform across individuals. For example, respondents may differ in the extent to which they are considered to be at risk of motivation, communication, memory or knowledge failures. Some of the cognitive question-answer processes described above assume that a respondent is sufficiently motivated to try to *optimise* their answers. However, Krosnick (1991) argues that when optimally answering a survey question *would require substantial cognitive effort*, some respondents might instead *satisfice* – i.e. provide what *appears* to be a satisfactory, rather than an optimal, answer. This could involve relying on some of the heuristics or biases listed in Table 2.1. Satisficing is particularly likely when the task is particularly difficult or for respondents with lower cognitive abilities, or those who might be tired, disinterested, impatient or distracted.

There is also evidence to suggest that a significant component of the variance in cross-sectional measures of self-reported survey data may be attributable to individual fixed effects – for example, individual differences in personality or generalised levels of affect, also known as "affectivity" (Diener, Oishi and Lucas, 2003; Diener, Suh, Lucas and Smith, 1999; Robinson et al., 2003; Suls and Martin, 2005). There also appear to be group differences in subjective well-being at the level of country (e.g. Vittersø, Biswas-Diener and Diener, 2005) and culture (e.g. Suh, Diener and Updegraff, 2008), which can operate over and above the differences that one might expect as a result of differences in current life circumstances. The extent to which these differences might be a result of differential susceptibility to error, or certain response styles, rather than substantive differences in underlying levels of subjective well-being, is hotly debated – and discussed in more detail in Section 5 of this chapter. Implications for data analysis are described in Chapter 4.

Summary – factors influencing patterns of error

Some of the factors thought to interact to influence the likelihood of error and/or biases in self-reported measures are summarised in Box 2.1. The essential question for this chapter is how potential sources of error and bias interact with – or can be caused by – aspects of survey methodology, when measuring subjective well-being.

Summary of issues investigated in this chapter

For ease of use, this chapter is organised around survey design and methodological considerations, ranging from the most specific (question construction) to the most general (the wider survey design and methods, and the role of response styles and cultural differences). Table 2.2 provides a guide to the key issues discussed and to some of the interactions or trade-offs between different elements of survey design that need to be considered.

1. Question construction

Introduction

The way a question is constructed influences respondent comprehension, information retrieval, judgement and reporting of subjective well-being. Seemingly small differences between questions may impact on the comparability of data collected through two or more different measures as well as their internal consistency and test-retest reliability over time. Questions that are easily understood, low in ambiguity, and not too burdensome for respondents should reduce error variability and enhance the validity of responses.

Box 2.1. **Factors thought to influence the likelihood of error, response biases and heuristics**

Factors associated with the underlying construct of interest	Survey design factors	Respondent factors
Task difficulty ● How easy or difficult is it for respondents to think about the construct or recall it from memory?	**Question wording** ● Is the wording complex or ambiguous? Can it be easily translated across languages and cultures? Is the tone of the question sufficiently neutral, or does it suggest particular answers should be favoured?	**Motivation** ● Are respondents equally motivated? **Fatigue** ● Are respondents equally alert and engaged?
Translatability ● How easy or difficult is it to translate the construct into different languages?	**Response formats** ● Is the wording complex, ambiguous or difficult to translate? Can the response options be easily remembered? Can respondents reliably distinguish between response categories? Are there enough response categories to enable views to be expressed fully?	**Susceptibility to social pressure, norms or demand characteristics** ● Do respondents vary in terms of their susceptibility to social pressure/or their likelihood of responding in a socially desirable manner?
Risk of social norms ● How likely is it that there are social norms associated with the construct, i.e. normatively "good" and "bad" answers?	**Question order** ● Do preceding questions influence how an item is interpreted and/or prime the use of certain information when responding?	**Language differences** ● Do language differences between respondents influence how respondents interpret questions and response formats?
Risk of influence by momentary mood ● How likely is it that respondents' momentary mood can influence how they remember/assess the construct of interest?	**Survey source/introductory text** ● Does the information provided to respondents suggest that a certain type of response is required (demand characteristics) or promote socially desirable responding?	**Cultural differences** ● Do cultural differences affect the type of response biases or heuristics that might be seen when respondents are satisficing?[1]
Risk of respondent discomfort ● How likely is it that respondents will find questions irritating or intrusive?	**Survey mode** ● Does the survey mode influence respondent motivation, response burden (e.g. memory burdens) and/or the likelihood of socially desirable responding?	**Knowledge** ● Do some respondents lack the knowledge or experience to be able to answer the question? (but attempt to do so anyway).
Respondent interest/engagement ● How relevant or interesting do respondents find the construct being measured?	**Wider survey context** ● Does the day of the week or the time of year affect responses? Could day-to-day events (such as major news stories) or the weather influence responses?	**Cognitive ability** ● Do respondents vary in their ability to understand the question and/or in their memory capacity?

1. Satisficing is when a respondent answers a question using the most easily available information rather than trying to recall the concept that the question is intended to address. A satisficing respondent may make use of a simple heuristic to answer the question or draw on information that is readily available in their mind rather than trying to provide a balanced response.

Question construction requires consideration of the precise wording of a question, and its translation into other languages where necessary, as well as the reference period that respondents are asked to consider when forming their answers (e.g. "in the last week" versus "in the last month"). As this topic is so large, detailed discussion of response formats are handled in their own separate Section 2, which follows – although of course the response formats that can be used will be influenced by how the initial question is framed.

Questions themselves are the most direct means through which a surveyor communicates their intent to respondents, but they are not the only source of information to which respondents may attend. Thus, comparability perhaps starts, but certainly does not end, with question construction – and other methodological factors influencing comparability are addressed in the sections that follow.

The section below describes issues of question construction in relation to evaluative measures, regarding assessments of life overall; affective measures capturing recent experiences of feelings and emotions; and psychological well-being or eudaimonic

Table 2.2. **Guide to the issues covered in this chapter**

Section of the chapter	Survey design issues under consideration	Key sources of error considered	Interactions between survey design issues
1. Question construction	• Question wording. • Length of the reference period.	Communication/translation failures. Memory failures. Response biases and heuristics.	Response formats (partly determined by question wording). Survey mode.
2. Response formats	• Number of response options to offer. • Labelling of response categories. • Unipolar versus bipolar measures. • Order of presentation of response categories.	Communication/translation failures. Memory failures. Response biases and heuristics.	Question wording (partly determines response format). Survey mode.
3. Question context, placement and order effects	• Question context and order effects. • Question order within a module of subjective well-being questions. • Survey source and introductory text.	Contextual cueing. Response biases and heuristics relating to demand characteristics, social desirability, consistency motif and priming effects.	Survey mode. Survey type (e.g. general household versus specific purpose).
4. Mode effects and survey context	• Survey mode. • When to conduct the survey.	Response biases and heuristics, particularly relating to respondent motivation/burden and social desirability. Contextual cueing as a result of wider survey context: • Day-to-day events. • Day of week. • Seasonal effects. • Weather effects.	Question construction and response formats. Survey type (e.g. general household versus specific purpose).
5. Response styles and the cultural context	• Risk of response styles. • Risk of cultural differences in response styles.	Consistent response biases and heuristics, associated with individual respondents. Cultural differences in characteristic response biases, heuristics and styles.	Cross-cutting section with relevance throughout.

measures. Some examples of these types of measures are included in Annex A. Many of the general principles discussed here will apply to other forms of self-report measures, but the emphasis here is on evidence relating specifically to subjective well-being, where this is available.

Question wording

The issue

In the case of subjective well-being measures, three key aspects of question wording are particularly important. The first issue is that of question comprehension. Establishing what a subjective self-report measure *actually means* for respondents is difficult – but for responses to be comparable, it is important that respondents interpret questions in a similar way.

Second, there is the issue of whether minor changes in question wording affect results in major ways. For the self-report survey method to be valid at all, it is essential that different question wording leads respondents to draw on different sources of information – so that, for example, asking respondents how happy they felt in the last 24 hours causes them to consider different information than asking them how angry they felt. Problems potentially arise, however, when different question wording is used to measure the *same* underlying construct. *How similar* question wording needs to be before it can be regarded as measuring the *same thing* is an essential issue for survey comparability.

The third issue is that of translatability – which includes both translatability between languages and cultures, as well as between generations and subgroups within a society. To minimise error variability, question wording needs to be understood in broadly the same way by different respondents. Certain question wordings may create particular difficulties when translated into different languages – or may have particular connotations for some groups of respondents. This potentially limits how comparable data are across different groups.

The evidence on question comprehension

It is difficult to evaluate the overall evidence on question comprehension, because different measures will perform differently. Question comprehension is usually examined through pre-testing and cognitive testing in the course of scale development, which is rarely reported in detail. Response latencies (i.e. the time respondents take to process the question, then construct and deliver their answers) can indicate how easily respondents can grasp a question and find the information needed to construct a response. Recent evidence (ONS, 2011a) indicates that 0-10 single-item questions on evaluative, eudaimonic and affective subjective well-being can be answered, on average, in around 30 seconds, suggesting that they pose no particular problems to respondents. However, short response latencies could also result if respondents are responding randomly or using mental short-cuts or *heuristics* to help them answer questions.

Empirical research on the validity of measures perhaps offers the best evidence that respondents have understood question meaning. If survey measures show a strong relationship with real-life behaviours and other non-survey indicators, this suggests that responses are meaningful. As described in Chapter 1, there is considerable evidence to support the overall validity of subjective well-being measures, particularly some of the most frequently-used life evaluation measures. However, this research typically draws on a wide range of different questions, and very few studies have systematically compared the impact of different question wordings under identical conditions.

The evidence on question wording

Evaluative measures. Several different questions have been used in an attempt to capture life evaluations – i.e. a reflective or cognitive assessment of life "as a whole". These include the Cantril "Ladder of Life" scale, general satisfaction-with-life questions and questions about global happiness (such as "Taken all together, would you say that you are very happy, pretty happy, or not too happy?"). Further illustrative examples of evaluative measures can be found in Annex A. These different measures are often described in the literature under the common umbrella terms "life satisfaction" or "happiness", but there is some evidence to suggest that different wordings may tap into slightly different underlying constructs. For example, Bjørnskov (2010) compared Cantril's Ladder of Life and a general Life Satisfaction measure across a range of countries and co-variates, finding considerable differences between these two measures.

Similarly, Diener et al. (2009) report evidence from more than 52 different countries indicating that, at the national level, responses to the Cantril Ladder are distributed somewhat differently – being centred more closely around the scale midpoint than other subjective well-being measures of life satisfaction, happiness and affect balance. Cummins (2003) cites evidence to suggest that the Ladder may produce more variable scores than standard life satisfaction questions – showing 2.5 times the degree of variation across eight different US population samples.

In contrast, Helliwell and Putnam (2004) examined the determinants of responses to both a global happiness and a life satisfaction question in a very large international data set ($N > 83\,500$), drawn from the World Values Survey, the US Benchmark Survey and a comparable Canadian survey. They found that, although the main pattern of results did not differ greatly between the two measures, the life satisfaction question showed a stronger relationship with a variety of social indicators (e.g. trust, unemployment) than did the happiness question. However, in this work happiness was measured on a four-point scale and life satisfaction was measured on a ten-point scale; it is thus not clear that question wording, rather than scale length, produced this difference.

In reality, it is often very difficult to separate the effects of question wording from other differences in survey design and administration, question presentation and response formats – and these are rarely investigated in a systematic fashion. One exception to this is recent experimental testing by the UK Office for National Statistics (ONS, 2011b). Using a common sample asked both a life satisfaction question and the Cantril Ladder question. The ONS showed that on the same 11-point scale, the mean average Cantril Ladder score (6.7) was lower than for life satisfaction (7.4), and the correlation between the two ($r = 0.65$) suggests that the responses to these questions were not identical.

More recently, in the *World Happiness Report* (Helliwell, Layard and Sachs, 2012), the authors systematically review how different evaluative questions perform in terms of country rankings, mean scores and co-variates using the Gallup World Poll, World Values Survey and European Social Survey. Questions on overall happiness, life satisfaction and the Cantril Ladder were found to produce essentially identical rankings ($r = 0.987$) and to have very similar co-variates, including the effect of income. When life satisfaction and the Cantril Ladder were asked of the same people in the Gallup World Poll, the correlation between the two measures was very high ($r = 0.94$). However, the Cantril Ladder did consistently produce a lower mean score than life satisfaction or overall happiness questions by about 0.5 on an 11-point scale across all the surveys considered.

Affect measures. In the case of affect measures, subtle forms of ambiguity can be introduced simply because respondents may not have an agreed definition or concept for a word. For example, the term "stressed" can have a variety of popular interpretations. Similarly, the term "anxious" could describe a clinically-significant condition of debilitating severity, a milder sense of unease and nervousness, or a sense of eagerly anticipating something (e.g. "I'm anxious to see what the film will be like"). In the absence of explicit definitions, respondents may seek contextual clues to infer intended meaning – for example, in other survey questions, or in their own knowledge and experiences.

The approach typically adopted in psychological research to reduce the impact of question ambiguity, or conceptual fuzziness, is to include multiple items in a scale – so that a broad construct, such as negative affect, is measured by a range of items (such as feeling *sad, upset, depressed, nervous, tense*, etc.), so that individual variability in how each of those items is interpreted may wash out when summing across all items. The improved reliability of multiple-item subjective well-being scales as compared with single-item measures (e.g. Schimmack and Oishi, 2005; Michalos and Kahlke, 2010) can thus potentially be attributed to their ability to reduce the impact of random error. For example, the UK ONS (2011b) reported that among a sample of 1 000 British adults, the overall distributions for 0-10

intensity measures of "anxious", "worry", and "stressed" were very similar, suggesting that even though different individuals may infer different meanings from individual items, population averages for each of the questions were quite similar.

Where only a limited number of questions are used to measure affect, the selection of which affective descriptors (e.g. *calm, happy, sad*) to include becomes very important. Item selection has been more rigorously examined in the affect literature than relative to the evaluative and eudaimonic measures, but there remains considerable variability across commonly-used measures in terms of which affective descriptors are included (Annex A contains examples). For the purposes of scale comparability, it will be important to ensure consistency across surveys in terms of the items used. As discussed in more detail below, selecting items or concepts that can be easily translated across languages will also be important for international comparability.

Eudaimonic measures. In the case of eudaimonia, several measures have been proposed – for example, Huppert and So's (2009; 2011) *Flourishing Scale*, and Diener and Biswas-Diener's (2009) *Psychological Well-Being Scale*, both of which are multiple-item measures, and the UK ONS (2011b) have also piloted a single-item eudaimonia question ("To what extent do you feel the things you do in your life are worthwhile?"). Details of each of these measures can be found in Annex A. The 14-item *Warwick-Edinburgh Mental Well-Being Scale* (e.g. Tennant et al., 2007), which asks respondents about feelings and thoughts experienced in the last two weeks, also contains items that are relevant to the construct of eudaimonia.

A more systematic comparison of these measures and their relative strengths in terms of both their validity and their usefulness to policy-makers is needed. The dimensionality of the construct will also be a crucial determinant of whether shorter or single-item measures can be developed for use in national surveys. Diener et al. (2009) investigated the 8-item *Psychological Well-Being Scale*, and found only one major factor (accounting for 50% of the variance in responses), on which factor loadings for individual items varied from 0.58 for feeling respected to 0.76 for leading a purposeful and meaningful life. However, Huppert and So (2011) reported a clear two-factor structure within their *Flourishing Scale*, the first being a cluster described as "positive characteristics" (including items concerning emotional stability, vitality, optimism, resilience, positive emotion and self-esteem), while the second reflected "positive functioning" (including engagement, competence, meaning and positive relationships).

The evidence on translatability

Although there is a very wide literature examining cross-cultural differences in subjective well-being, there is little in the published literature that specifically reports on the influence of translation. What we do know in this field raises some concern – for example, in the case of evaluative measures, Bjørnskov (2010) states that the English word *happy* is "notoriously difficult to translate", whereas the concept of *satisfaction* better lends itself to precise translation (p. 44). Veenhoven (2008) equally notes that perfect translation is often not possible: "If it is true that the French are more choosy about how they use the word "happy", they might place the option "'*très heureux*' in the range 10 to 9, whereas the English raters would place "very happy' on the range of 10 to 8" (p. 49). When Veenhoven tested this issue with Dutch and English students, the Dutch rated "very happy" as being equivalent to 9.03 on a 10-point scale, whereas the English rated it at 8.6 on average.

Factor analysis is often used in selecting the most appropriate wording for multiple-item scales to test whether all items appear to be measuring the same underlying construct. For example, Thompson (2007) used a combination of bilingual focus groups and factor analysis to develop a short version of the Positive and Negative Affect Schedule. On the basis of evidence gathered from 12 different nationalities, the words *enthusiastic, strong, interested, excited, proud, scared, guilty, jittery, irritable* and *distressed* were rejected from the final measure, either because they had ambiguous meaning for some respondents (for example, both positive and negative connotations) and/or because they performed poorly in terms of factor structure. Ultimately, the words *alert, inspired, determined, attentive* and *active* were selected to measure positive affect, and the words *upset, hostile, ashamed, nervous* and *afraid* were selected to measure negative affect.

Although precise question wording is usually helpful for communicating concepts to respondents, broad but easily-translatable descriptors may be particularly helpful to produce internationally-valid affect measures, where some emotion words may be too narrow or culturally-specific to generalise across samples. For example, Diener et al. (2009) make a case for including very broad descriptors for positive and negative feelings in affective experience measures to avoid the potential omission of feelings that are not easily translated, particularly those that might be more important in some cultures or to certain individuals. As a result, their 12-item Scale of Positive and Negative Experience (SPANE) includes quite generic items about feeling positive, negative, good, bad and pleasant and unpleasant.

Key messages on question wording

Using different wording in subjective well-being questions can change the pattern of responses, although finding a difference between response patterns does not in itself indicate which wording should be preferred. The evidence regarding the "best" question wording to use is limited and mixed; indeed, as discussed in the section that follows, there are a wide variety of question and survey context features that can influence responses in a way that cannot be easily disentangled from that of wording effects alone. There are also some grounds for concern about the translation of certain subjective well-being constructs between languages, although the size of this problem and the extent to which it limits cross-national comparability is not yet clear.

One way to reduce the impact of potential variation in how respondents understand questions is to use multiple-item scales, which approach the construct of interest from several different angles in the hope that they ultimately converge. Current measures of affect and eudaimonia typically contain multiple items – which also enables conceptual multi-dimensionality to be examined. In a national survey context, lengthy multiple item scales may not be practical, but current evidence suggests that particular care needs to be taken when developing shorter or single-item affect and eudaimonia measures, as there is strong evidence for multi-dimensionality among these measures. Although multiple-item life evaluation measures show better reliability than their single-item counterparts, there is at present evidence to suggest that single-item measures *can* be successfully used to capture life evaluations, which are usually assumed to be unidimensional in nature. It would be useful, however, to have further evidence regarding the relative accuracy and validity of single- versus multiple-item evaluative questions to help ensure optimal management of the trade-off between survey length and data quality.

Psychometric analysis, including examination of factor structure and scale reliabilities, can go some way towards identifying questions that seem to function well among a set of scale items, but consideration of the construct of interest, validation studies and the *purpose* of measurement will also determine which question should be selected. For now, this discussion clearly highlights the need for a standardised approach to question wording, particularly if comparisons over time or between groups are of interest.

Length of the reference period

The issue

Subjective well-being questions often ask respondents to sum their experiences over a given reference period – such as emotions experienced *yesterday*, or satisfaction with life *nowadays*. Selection of the reference period ultimately needs to focus on the purpose of the measure (i.e. the underlying construct of interest) – and, particularly in the case of affect, the reference period plays a key part in defining *what* is being measured (because affect yesterday and affect last year are different things).

There are two central ways in which the reference period can influence the comparability of responses and the risk of error. The reference period provides information to respondents about the construct of interest (e.g. a measure of *happiness* experienced over one year might tap global life evaluations, whereas *happiness* experienced over one day is likely to capture more short-term affect). Second, if the reference period is too demanding – for example, if respondents are asked to recall over too long a time period – it may lead to misremembering and increase susceptibility to recall and response biases and context effects.

These issues raise two further questions. First, how similar do two reference periods need to be for measures to be considered comparable? And second, given the potential risks for error, what is the optimal reference period for each type of subjective well-being measure?

The evidence – evaluative measures

Rather than specifying a particular reference period, evaluative questions typically ask respondents how their life is overall *nowadays* or *these days* (examples at Annex A). Thus, whilst inviting an assessment of life that might span the entire life-course, respondents are asked to evaluate their life *at this point in time*, which is likely to favour the recent past (Diener, Inglehart and Tay, 2012). Cognitive testing work undertaken by the UK's ONS in the course of developing experimental questions on subjective well-being (ONS, 2011c) indicated that "nowadays" was interpreted by respondents in several different ways, ranging from the present moment, the last five years, or since key life events. "Nowadays" was also considered to be a dated or old-fashioned term. The term "these days" was meanwhile considered as referring to a shorter time-frame (e.g. the past few weeks).

It is difficult to identify definitively a right or wrong reference period for evaluative measures. As Diener, Inglehart and Tay (2012) note, there has thus far been little research into i) how the framing of life satisfaction questions influence scores; and ii) which framing is most valuable for policy. For data to be comparable, consistency in the use (or not) of reference periods is certainly preferable – although it is not at present clear how important this consistency is, and how precisely reference periods need to match. Some of the most widely-used evaluative measures in the literature at the moment ask *at present* (Gallup World Poll) and *these days* (World Values Survey) – but it seems unlikely that these two

terms should produce large differences in responses. Greater variability may be found between those questions that use a reference period and those that do not, although again the evidence regarding the precise size of the impact this may have is lacking.

The evidence – affect measures

The length of reference period is likely to be of particular interest in relation to measures of recently experienced affect, where existing methods range from asking respondents about the intensity of feelings *right now* (e.g. in the experience sampling method, where respondents often report on their affect levels several times a day) to the presence or absence of emotions *yesterday* (e.g. the Gallup World Poll) or the frequency of positive and negative experiences over the *past four weeks* (e.g. the Diener and Biswas-Diener *Scale of Positive and Negative Experience*, 2009). Some examples of existing affect measures are provided in Annex A. Of course, to the extent that these measures are designed to capture *different constructs*, such as momentary mood versus more stable patterns of experience, differences in responses can be due to valid variance. The issue in these cases is to select the best measure for the construct of interest. However, the extent to which approaches differ also affects both the error sources and the comparability of the data, and thus needs to be considered when selecting measures and interpreting the findings.

The importance of reference periods to affect measures has been examined by Winkielman, Knäuper and Schwarz (1998), who asked respondents how frequently they *get angry* within a 1-week reference period, and within a 1-year period. Consistent with the hypothesis that shorter reference periods prompt respondents to report more frequent experiences, and longer reference periods are assumed by respondents to pertain to rarer experiences, they found that fewer but more extreme episodes were reported within the 1-year period. Asking *how often* an individual experiences anger also presupposes that individuals *will have* experienced this emotion within the given timeframe, whereas prefixing this with a yes/no "were you ever angry during the last week…" question would send quite a different signal about the expected frequency of an experience. Arguably, the phrase *get angry* is simply too vague to have independent meaning – and thus the time frame associated with it provides respondents with information about how it should be interpreted. If the key concept of interest is better defined, the reference period may play less of a role in understanding meaning. However, particularly in the case of affect, the constructs of interest are intrinsically linked to the reference period – such that *affect yesterday* is a fundamentally different construct to affect in the last four weeks.

Whilst respondents might be expected to provide reasonably accurate recall of emotional experiences in end-of-day reports (e.g. Parkinson, Briner, Reynolds and Totterdell, 1995; Stone, 1995), quantifying the extent to which an emotion was experienced in terms of either frequency or intensity over a longer period may be more challenging. Cognitive testing by the UK's ONS (2011c) indicated that some respondents find remembering their activates on the previous day quite difficult, and thus averaging the intensity of emotions over even just one day may be quite cognitively demanding. However, the ONS also reported that some respondents objected to questions that focus on affective experiences on a single day, raising concerns that this may be unrepresentative of how they usually feel.[2] This has implications for both respondent motivation and accuracy of reporting, and suggests that interviewer briefing and the preamble provided for short-term affect questions will be particularly important (discussed in Chapter 3).

To better understand accuracy of recall for affect, Thomas and Diener (1990) compared small-scale experience sampling (Study 1) and daily diary (Study 2) data with retrospective accounts of mood from the same respondents across a period of three and six weeks respectively. In both studies, they found that retrospective reports of affect intensity were significantly higher than actual intensities experienced (for both positive and negative affect). Frequency judgements regarding the proportion (0-100%) of time when positive affect was stronger than negative affect proved to be more accurate, but the estimate was in this case significantly *lower* than the actual frequency. Perhaps most compelling, however, was evidence from the cross-individual correlations between retrospective reports and actual experiences. Although several of these correlations were highly significant, they suggest a considerable gap between experiences and retrospective reports.[3]

These findings, although limited, suggest that retrospective assessments over several weeks are likely to produce lower quality data than those over just 24 hours. Asking respondents to report on their affective experiences over long time-frames may still reveal something interesting about their overall level of subjective well-being (e.g. by tapping into dispositional components of affect), but may not be reliable as an accurate account of *experienced* affect or emotion. There is also tentative evidence that asking respondents to recall the *frequency* of their emotional experiences may produce more accurate reports than *intensity* judgements.

The evidence – eudaimonic measures

Most eudaimonic measures do not tend to specify a reference period for respondents. For example, neither the eudaimonia nor *flourishing* scale created from items in the European Social Survey (Huppert et al., 2009; Huppert and So, 2011; Clark and Senik, 2011) nor the *Psychological Well-Being Scale* proposed by Diener and Biswas-Diener (2009) provide specific guidance to respondents about the reference period in question. The European Social Survey refers to what respondents *generally* or *always* feel, whereas the Psychological Well-Being Scale asks respondents the extent to which they agree or disagree with a series of statements about their lives (e.g. *I lead a purposeful and meaningful life*) – further examples are available at Annex A. Respondents are thus presumably expected to indicate their views at the present moment, but looking back across their lives for an undefined period. As with evaluative measures, it is not yet clear to what extent responses might be altered by different reference periods. As a consequence, it will be important to maintain a consistency of approach until further evidence is available.

Key messages on the length of the reference period

Reference period can have a strong impact on affect measures – and indeed, partly defines *what* is being measured. As a consequence, it is vital to consider the ultimate purpose of the measure when selecting the reference period. If short-term affective experiences are the variable of interest, the evidence generally suggests that reports within 24 hours are needed. Longer reference periods may meanwhile pick up dispositional affective tendencies – although the retrospective recall burden is also expected to produce greater measurement error.

Less is known about the impact of reference period on evaluative and eudaimonic measures, which tend to be less specific about timing. Reference period may be particularly important when researchers are tempted to modify subjective well-being measures for specific uses – for example, to evaluate the impact of a given policy intervention over a

specific time period. However, for more general questions, the widely used approach of focusing on "these days" or "at present" may be advised for evaluative measures. This remains an area where further research would be valuable.

2. Response formats

Introduction

Although providing respondents with a rating scale may seem straightforward, there are many ways in which response formats can vary. Units are an integral part of measurement – and in surveys, the response format both specifies the units of measurement and the possible range of those units. Question designers must make decisions about how many response options to offer, how to label those response options, and whether and how to label scale intervals, as well as taking some more fundamental decisions on whether a particular aspect of subjective well-being should be measured on a bipolar scale (e.g. *agree/disagree*) or a unipolar scale (e.g. *not at all-completely*), and whether respondents should be asked for a judgement involving frequency (*how often do you feel…?*) or intensity (*how much do you feel…?*).

Response formats matter – for the validity, reliability and comparability of responses. Choosing the "right" response format means choosing a format that adequately represents the construct of interest (e.g. is the focus on the direction of feeling, or on both its direction and intensity?). It also means selecting a format that respondents can understand and use accurately, effectively and consistently, i.e. one that has the same meaning for all respondents, under a variety of circumstances. The optimal format also needs to capture the full range of response possibilities adequately and to allow respondents to express where they perceive themselves within this range. For example, asking respondents whether they watch 1, 3 or 7 hours of TV per day, or whether they feel either entirely delighted or completely terrible, would not enable some respondents to give a valid and meaningful answer.

Finally, there may be differences in the response formats that may be optimal for evaluative, eudaimonic and affective measures. Evaluative and eudaimonic measures are similar to attitude measures in that it may be preferable for the response format to contain information about both the direction of feeling (positive/neutral/negative or agree/disagree), as well as its intensity (strong-weak). In the case of affect measures, it is often desirable to measure positive and negative affective states separately. Thus, rather than asking about the direction (positive-neutral-negative) of affect, respondents are often given a single adjective (e.g. *happy*) and asked to describe either the intensity or the frequency with which they felt that way within a given time period. This may in turn have implications for the optimal number of response options, as well as response scale labelling and anchoring.

The number of response options to offer

The issue

Sensitivity or discriminatory power is an important factor in scale design. A good subjective well-being measure will be one that enables variability between respondents to be detected where it exists. In addition to the way in which a question is worded, the number of response options offered is a critical determinant of scale sensitivity – and this is particularly so for measures that rely on just one question, rather than summing across a number of items.

Offering too few response options may not enable some respondents to fully express themselves. This can cause meaningful variations in scores to be lost, and respondents may also grow frustrated, leading to lower quality or even random responding. On the other hand, too many response categories can potentially increase cognitive burdens, especially if the response options presented offer finer distinctions than respondents are able to make in their own judgments or on the basis of recall from memory. Thus, offering too many response options also has the potential to demotivate respondents, leading to lower-quality data.

There is, therefore, a trade-off between capturing as much meaningful variation as possible on the one hand, and minimising respondent burden and frustration on the other. Survey mode also has important consequences for the cognitive burden of different response formats, something that will be discussed in more detail later. The preferred number of response options may also be different for different types of subjective well-being measures, given the differences in the underlying constructs. Finally, respondents may vary in their ability to cope with cognitive burdens and in their preferences for simple versus complex response options. Thus a compromise must be struck between the needs of the data users, the respondents being surveyed, and the constraints of the survey method.

The evidence – general research on the number of response options

Whilst offering a large number of response options will not automatically lead to greater discriminatory power,[4] offering too few response categories precludes being able to detect finer variations among respondents where these do exist. From this one perspective, then, a longer scale with more response options is better. There is also a range of evidence to suggest that, among attitude measures, longer numerical response scales (i.e. numerical scales with a range of numerically-labelled response options, but verbally-labelled anchors at the scale extremes) tend to increase both internal consistency and test-retest reliability, although the gains do not appear to be large (Weng, 2004; Preston and Colman, 2000; Alwin and Krosnick, 1991). Finally, there is some evidence from consumer research to suggest that validity increases with increasing numbers of response categories or scale points (Preston and Colman, 2000), again with numerical scales.

There is, however, considerable debate around the *optimal* number of response categories – and a very wide range of opinions is available in the literature (see Weng, 2004, for a brief summary). This number will depend on respondents' information-processing capacities and preferences, survey mode, scale labelling, and, to some extent, presentational concerns and questionnaire length. Increasing the number of response categories beyond the optimal length could result in loss of information, increased error and decreased reliability, because the individual scale points will mean less to respondents. The increased response burden associated with longer scales may also lead respondents to become less motivated to optimise and more likely to satisfice in their answers, thus also increasing the risk of response biases and error (Alwin and Krosnick, 1991; Alwin, 1997).

So, how many response categories is enough, and how many is too many? In scales where all response categories are given verbal labels, Bradburn et al. (2004) argue that, due to the burden on memory and attention, five categories is the maximum number that a respondent can process in a verbal interview setting (telephone or face-to-face) without visual prompts. Furthermore, when the response categories are qualitatively different from one another (rather than being imagined on a sliding scale), these authors suggest that

four categories should be the upper maximum. On the other hand, Alwin and Krosnick (1991) indicate that respondents may prefer to have response options denoting weak, moderate and strong negative and positive evaluations (i.e. a 7-point scale) in part because these are the categories that people often use to describe attitudes and opinions in everyday life.

For numerical scales, which might be anchored by descriptive adjectives at the two scale extremes, it is easier for respondents to attend to, and respond successfully to, longer scales, because only the scale end-points need to be retained in memory. The limiting factor in these cases becomes more an issue of whether respondents can reliably discriminate between categories. Bradburn et al. (2004) state that while sensory research traditionally uses 9-point scales, psychometric evidence indicates that most respondents cannot reliably distinguish between more than six or seven levels of response. It is notable how little of the literature in this field makes explicit reference to scale *content* when describing optimal scale length – as though there were fundamental limits on the human ability to discriminate between response options regardless of what those options refer to.

Another consideration when attempting to determine the number of response categories to use is whether or not to include a scale mid-point. Chang (1994) highlights previous work indicating that scales with an odd number of points (and therefore a natural mid-point) can result in respondents selecting the middle category by default – and there is reduced utility in the measure if large numbers of respondents select the same response category, as discriminating power will be diminished. An even number of response categories, on the other hand, typically forces respondents to express a directional preference for one of the scale anchors over the other, even where they may not have such a preference. In the measurement of bipolar attitudes (e.g. disagree/agree), Alwin and Krosnick (1991) and Bradburn et al. (2004) argue in favour of including a scale mid-point (and therefore an odd number of response categories) to enable respondents to express a neutral position and to prevent random responses from those with no attitude or with genuinely ambivalent feelings.

One final but important consideration is the scale length that respondents seem to prefer. As noted earlier, the optimal number of response categories is bounded by expressive capacity at the lower end (i.e. what is the minimum number of response options needed for respondents to feel like they can represent their experiences accurately?) and by processing capacity at the upper end (i.e. how many response options can individuals reliably discriminate between, and how many can respondents hold in memory or simultaneously compare?). Although respondent preferences perhaps provide only one perspective on these issues, they may offer important insights into the accuracy with which individuals are *actually able to use* the scale, and they may have important consequences for motivation, which in turn influences the quality of the data obtained.

There appears to be little empirical literature examining respondents' views on scale length – although it is the kind of question that (typically unpublished) cognitive testing may have addressed. In the context of a customer services questionnaire,[5] Preston and Colman (2000) tested a variety of scale lengths, from between 2 to 11 response options, as well as a 101-point scale. For "ease of use", scales with five, seven and 10 response categories were the most preferred, and the scales with 11 and 101 response categories were least preferred. Shorter scales were regarded as the most "quick to use", and again scales with 11 and 101 categories fared the worst. However, when asked which scale

"allowed you to express your feelings adequately", the two-point and three-point scales received extremely low ratings, and scales with 9, 10, 11 and 101 response categories were rated highest.

Preston and Colman's findings differ, however, from another marketing study by Dolnicar and Grün (2009), which also used pen-and-paper questionnaires, this time to examine attitudes and behavioural intentions with regards to environmental issues. Respondents were asked to evaluate a binary (*disagree/agree*), ordinal (7-point Likert scale ranging from *strongly disagree* to *strongly agree*) and a "metric" answer format (where respondents marked their position on a dotted line anchored by the words *strongly disagree* and *strongly agree*). There was no difference between response formats in terms of the perceived simplicity of the scale, or the extent to which respondents felt they could express their feelings. The binary format was, however, perceived to be quicker to use – and this was supported by actual measures of time taken to complete the surveys, which was significantly faster in the case of the binary measures. Dolnicar and Grün also observed specific patterns of number use on ordinal and metric response formats: in both cases, respondents showed a tendency to use more extreme answer options when asked about behavioural intentions, and more moderate answer categories when asked about beliefs. No differences were observed in how respondents used binary scales when answering these different types of questions.

The evidence – evaluative measures

Life evaluations are often assessed through a single question, which means the number of response options offered is a particularly important determinant of scale sensitivity (discriminating power). Cummins (2003) recommends that 3-point response scales should be eliminated from life satisfaction research because they are too crude to be useful in detecting variation in responses. As Bradburn, Sudman and Wansink (2004) point out, on a single-item scale with three response categories, anchored by extremes at either end (for example, "best ever", "worst ever" and "somewhere in between") most people will tend to select the middle category. Alwin (1997) meanwhile argues that, if one is interested in attitudes that have a direction, intensity and region of neutrality, then a minimum of five response categories is necessary (three-category response formats communicate neutrality and direction, but not intensity).

Increasing the number of response options available is unlikely to make a difference unless the options added represent meaningful categories that respondents consider relevant to them. Smith (1979) examined time-trends in US national data on overall happiness and reported that extending the scale from 3 to 4 response options through the addition of a *not at all happy* category captured only a small number of respondents (1.3%), and did not lower the mean of responses (the *not very happy* category simply appeared to splinter). Conversely, adding a fifth *completely happy* response category at the other end of the scale both captured 13.8% of respondents and drew respondents further up the scale, with more respondents shifting their answers from *pretty happy* to *very happy*.

For evaluative measures with numerical response scales, longer scales (up to around 11 scale points) often appear to perform better. Cummins (2003) argues that there is broad consensus that a 5-point scale is inferior to a 7-point scale. Alwin (1997) compared 7-point and 11-point scales on multi-item measures of life satisfaction. Using a multi-trait-multi-method design, Alwin found that across all 17 domains of life satisfaction measured, the 11-point scales had higher reliabilities than the 7-point scales. In 14 out of 17 cases, the 11-point scales

also had higher validity coefficients; and in 12 of 17 cases, 11-point scales had lower invalidity coefficients, indicating they were affected less, rather than more, by method variance – i.e. systematic response biases or styles. This overall finding is supported by Saris et al. (1998) who used a similar multi-trait-multi-method analysis to compare 100-point, 4 or 5-point and 10-point satisfaction measures, and found that the 10-point scale demonstrated the best reliability. Similarly, in a German Socio-Economic Panel Study pre-test, Kroh (2006) also found evidence that 11-point satisfaction scales had higher validity estimates than 7-point and open-ended magnitude satisfaction scales.[6]

Where it is desirable to make direct comparisons between measures (e.g. for international analysis, or between differently-worded questions), it will be important that measures adopt the same number of response options. Although procedures exist whereby responses can, in theory, be mathematically re-scaled for the purposes of comparison, there is evidence to suggest that, when life evaluation data are re-scaled in this way, longer scales can produce higher overall scores, thus potentially biasing scores upwards. For example, among the same group of respondents measured at the same time point, Lim (2008) found that the mean average level on an 11-point scale of global happiness[7] was significantly higher than recalibrated 4- and 7-point measures (although, curiously, not the 5-point measure). Lim attributed this to the negative skewness typically observed in the distribution of evaluative measures of happiness and life satisfaction. Cummins (2003) reports a similar finding, and further argues that this negative skewness means that life satisfaction measures are perhaps *best* scaled with at least 11 points, as most of the meaningful variance is located in the upper half of the distribution.

The evidence – affect and eudaimonia measures

Although there is an emerging trend towards 11-point scales for life evaluations, in the affect literature many authors still rely on 5-point measures, particularly in the case of experienced affect (e.g. Diener et al., 2009). Eudaimonia scales also tend to have fewer response categories (typically 5 or 7). Because the majority of existing affect and eudaimonia measures contain multiple items in order to assess each hypothesised underlying construct (e.g. 5 items to measure positive affect; 5 items to measure negative affect) and responses are then summed across a variety of items, overall scale sensitivity may be less severely threatened by a smaller number of response options, relative to single-item measures. But more work is needed to examine this further.

Another possible reason for the predominance of shorter scales in this literature may be that, while it might be relatively straightforward to assign reasonable verbal labels to a 5-point scale (e.g. *never, rarely, sometimes, often, always; strongly agree, agree, neither agree nor disagree, agree, strongly agree*), devising seven or more verbal categories strays into vague terms that do not have a clearly accepted position (e.g. *quite often; slightly agree*). One obvious solution to this challenge is to adopt numerical scales (e.g. 0-10) where only the scale end-points are labelled (e.g. from "never" to "all the time" or "not at all" to "completely"); this is the approach that has been adopted by the UK's ONS (2011b) in their experimental measures of subjective well-being. Further development of such scales could helpfully examine whether respondents are actually able to make meaningful discriminations between eleven different response categories when it comes to reporting affective experiences and eudaimonia.

Key messages on scale length

The optimal number of response categories in rating scales will be informed by a combination of reliability, validity, discriminating power and respondent preferences. Some consideration also needs to be given to enabling comparisons with the existing literature, which includes tried-and-tested response formats. On balance, the evidence broadly seems to favour an 11-point numerical scale for evaluative subjective well-being measures – and this is consistent with some of the current most widely-used approaches (including the Gallup World Poll and the German Socio-Economic Panel). Less is known about the optimal number of response options to offer for eudaimonic and affective measures, and this needs to be considered in combination with other factors that influence response formats, such as survey mode, and whether verbal or numerical scale labels are preferred (discussed below).

It is surprising how little reference the literature makes to the underlying construct of interest when discussing the optimal number of response options to present – even though the initial evidence suggests that this could be important. Life evaluations and eudaimonia measures are often presented to respondents in the form of "bipolar" attitude measures (e.g. *completely dissatisfied* to *completely satisfied* or *disagree completely* to *agree completely*). The intent behind this is to capture both the direction (negative-positive) and intensity (strong-weak) of feeling. 11-point numerical scales appear to be well-suited to this task, and offer the additional advantage of a scale midpoint, which provides respondents with the option of indicating they fall somewhere in between the two scale end-points, rather than forcing a choice in favour of one side or the other. For single-item scales, including some of the most frequently-used life evaluation measures, 11-point numerical scales also perhaps offer the best balance between scale sensitivity and so much choice that respondents are overwhelmed.

For affect measures, one might be interested in measuring either the intensity of feeling or the frequency with which that feeling occurred. Measures of recently-experienced affect are less like attitude measures, in that one is effectively asking respondents to remember a specific experience or to sum experiences over a specific time period. Affect measures also differ from many evaluative and eudaimonic measures in that they may not be seeking information about the *direction* of feeling, because it is desirable to measure positive and negative affective states separately (see the section on unipolar versus bipolar measures, below). Finally, affect measures are typically assessed through multi-item measures, which means that scale sensitivity is less strongly determined by any single item. However, there appears to be a lack of research that systematically examines these factors in combination with one another.

Scale labelling

The issue

Similar to reference periods, the way in which response formats are described or labelled sets a reference frame for individuals as they construct their answers. Scale labels can thus potentially influence mean values by communicating information about the *expected range* within which responses can fall. Variations in labelling the response scale can also affect the accuracy and reliability of scores, with response options that lack meaning or clarity being more likely to introduce random responding or satisficing.

A key decision for question design is whether to provide a verbal label for every response option (e.g. would you say that you are *very happy, pretty happy*, or *not too happy*?) or whether to simply label the scale end-points or anchors (e.g. 0 = *completely dissatisfied*; 10 = *completely satisfied*) and rely on numerical labels for scale intervals. As highlighted in the previous section, this choice will need to be considered in combination with other factors, such as the number of response options to offer and the survey mode being used.

The evidence – labelling scale anchors

Scale anchors (i.e. how scale end-points are labelled) send a signal about the range within which responses are expected to fall. This can be particularly important when the concept in question is vague, difficult to answer, or difficult for the respondent to apply to their own situation. However, even in the case of more objective behavioural reports, scale anchors can influence the overall distribution of responses, so that for example a scale anchored by *several times a day* at one end and *twice a month or less* at the other can elicit higher frequency self-reports than a scale anchored by *more than twice a month* at one end and *never* at the other (Schwartz et al., 1985; Schwartz, 1999; Wright, Gaskell and O'Muircheartaigh, 1994). Schwartz and colleagues (e.g. Schwartz, 1999; Schwarz, Knäuper, Oyserman and Stich, 2008) also provide evidence suggesting that respondents assume that values in the middle of a scale reflect the *average*, whereas the ends of a scale reflect distributional extremes. One approach suggested by Diener et al. (2009), which capitalises on this tendency, is to anchor the response scale with absolutes (for example, *always* and *never*), as these should in theory have the same meaning to all respondents, and make clear that all possible variations in between are captured. Of course, this general advice needs to be considered in the context of the underlying construct of interest, and may be easier to apply in some cases than in others.

There is some evidence to suggest that scale anchors may also affect the prevalence of certain response biases. For example, according to Krosnick (1999) the use of *agree/disagree*, *true/false* and, to a lesser extent, *yes/no* scale anchors can be problematic because they are more susceptible to acquiescence bias or "yea-saying", i.e. the tendency to endorse statements regardless of their content. This does not appear to have been well-tested with subjective well-being measures, despite the fact that eudaimonia measures in particular frequently adopt the *agree/disagree* format. Green, Goldman and Salovey (1993) have, however, suggested that a *yes/no* checklist-style response scale can produce greater positive endorsement of affect items. Whilst all self-report scales may be at risk of acquiescence if respondents are fatigued, unmotivated or overburdened, acquiescence may also interact with social desirability in the case of subjective well-being, due to the positive social value placed on eudaimonia, life satisfaction and positive affect in general.

The evidence – labelling scale intervals or response options

A key choice in designing questions is how to label scale response options or intervals. For most measures of subjective well-being there are no objectively identifiable scale intervals – that is to say, there are no pre-defined "levels" of *satisfaction*, or *happiness*, etc. Despite this, many of the first life evaluation measures used in the 1970s asked respondents to distinguish between responses such as *very happy, fairly happy* and *not too happy*. An alternative approach, adopted in many of the more recent life evaluation measures, such as the later editions of the World Values Survey and the Gallup World Poll, asks respondents to place themselves somewhere along a 0-10 numerical scale, where only the scale anchors are given verbal labels. In the case of recently experienced affect, frequency scales are more

commonly used (e.g. ranging from *never* to *always*), although numerical scales and simple *yes/no* judgements are also used. In the case of eudaimonia, the *agree/disagree* response format is often adopted, as with many attitude measures used in polling. Annex A provides illustrative examples of the range of scale labels used in the literature across the different types of subjective well-being measures.

It is clear that consistency in scale labelling is important for scale comparability. There is evidence that even subtle differences in scale labels can have notable effects on the distribution of subjective well-being scores. Examining time-trends in US national happiness data, Smith (1979) observed that a change in the wording of response categories caused shifts in patterns of responding. For example, on a three-item scale, offering the response category *fairly happy* instead of *pretty happy* was associated with more respondents (around 1.5 times as many) selecting the next response up the scale, *very happy*. This implies that *fairly happy* was perceived less positively than *pretty happy*. Similarly, the response options *not happy* and *not very happy* seemed to be perceived more negatively than *not too happy*, which attracted around 3.5 times as many respondents as the former two categories.

There is some evidence to suggest that providing verbal labels for response categories along numerical scales may influence the distribution of responses. For example, Pudney (2010) found that the labelling of response scales had a significant impact on the distribution of responses across a range of different satisfaction domains, although this finding was significant only for women, and was weaker in the cases of income and health satisfaction. Specifically, labelling only the scale anchors tended to reduce the mean level of reported satisfaction relative to adding verbal labels for all scale points. In further multivariate analyses, however, differences in scale labelling did not produce systematic or significant differences in the relationships between various predictors (e.g. health, income, work hours, etc.) and the satisfaction measures. So, although the differences in distributions produced by different scale labelling are of concern, they may not have very large impacts on the relationships observed between measures of satisfaction and its determinants.

Several authors have suggested that it is optimal to provide verbal labels for *all* numerical response options. Conti and Pudney (2011) analysed the impact of a change in response labelling on the job satisfaction question included in the British Household Panel Survey (BHPS) between 1991 and 1992 survey waves. They reported that providing verbal labels for some, but not all, response categories could draw respondents to favour the labelled categories. This work highlights the importance of examining not just the mean scores but the *distribution* of scores across the different response categories. However, one factor not discussed by Conti and Pudney is the impact of a change in one of the scale anchors between survey waves.[8] Thus, the change in distribution of scores between 1991 and subsequent years could be a product of a change in the scale anchor, the addition of verbal labels or a combination of the two features.

It has been suggested that adding verbal labels to all numbered response options can help clarify their meaning and produce more stable responding (Alwin and Krosnick, 1991). This was supported by Alwin and Krosnick's analysis of the reliability of political attitude measures over three waves of five different sets of national US panel data. The adjusted mean reliability for fully labelled 7-point scales was 0.78, whereas for numerical 7-point scales with only the endpoints labelled, this dropped to 0.57, a significant difference. Although much less dramatic than Alwin and Krosnick's finding, Weng (2004) provides

further evidence among a sample of 1 247 college students that textual labelling of every response category can increase test-retest reliability on 7- and 8-point scales (but not for 3-, 4-, 5- and 6-point scales).

Although the studies cited above generally imply that full verbal labelling of all scale intervals is preferable, and that adding verbal labels to response categories can have a significant impact on the distribution of responses, none provides conclusive evidence that full verbal labels offer a clear improvement in terms of *scale accuracy* or *validity*, and there is some evidence (Newstead and Arnold, 1989) that numerical ratings may be more accurate than labelled scales. In terms of discriminatory power, full verbal labels on the satisfaction measures examined by Pudney and by Conti and Pudney actually produced a heaping of responses on one response category (*mostly satisfied*) and appeared to increase the skewness of the data, which could be unhelpful in analysis based on linear regression models. A further practical difficulty is that adding verbal labels to a scale with nine, seven or possibly even only five response categories will make it challenging for respondents to answer in telephone surveys (without visual prompts) due to the memory burden involved. This could in turn limit the quality of the resulting data and further reduce comparability between different survey modes.

One of the challenges of using verbal scale labels, however, is that when only the vague verbal labels are used to denote intervals on a scale, it is not possible to know whether the categories are understood in the same way by all respondents. Several authors indicate that the use of vague quantifiers (such as *a lot, slightly, quite a bit* or *very*) should be avoided, as these can be subject to both individual and group differences in interpretation (Wright, Gaskell and O'Muircheartaigh, 1994; Schaeffer, 1991). Schwarz (1999) describes vague quantifiers as the "worst possible choice", pointing to the fact that they are highly domain-specific. For example, "frequently" suffering from headaches reflects higher absolute frequencies than "frequently" suffering from heart attacks" (Schwarz, 1999, p. 99). It has been suggested that numerical scales can help to convey scale regularity (Maggino, 2009) as they are more likely to be interpreted by respondents as having equally spaced intervals (Bradburn et al., 2004), although empirical evidence is needed to support this. The optimal way to label scale intervals also strongly interacts with survey mode, the number of response options presented, and the number of questions (items) used to measure each construct. One key advantage in terms of scale sensitivity is that numerical scales appear to enable a wider range of response options to be offered (because respondents only need to hold the verbal descriptions of the scale anchors in memory, rather than the label for every response option). As noted above, it has been suggested that, particularly for telephone interviews (where show cards and visual prompts are less likely to be used), only around four verbal response options can be presented before respondents become over-burdened. For measures involving just a single question, this can place severe constraints on scale sensitivity.

Verbally-labelled response scales may also present particular translation challenges. There may be both linguistic and cultural differences in how verbal labels are interpreted – and Veenhoven (2008) presents evidence to suggest that English and Dutch respondents assign different numerical values to the labels *very happy, quite happy, not very happy* and *not at all happy*.

Key messages on scale labelling

A number of complex trade-offs need to be managed when making decisions about how to label response options – including interactions with the overall number of response options, the survey mode and the translatability of items – as well as, of course, the underlying construct of interest and the manner in which the question is phrased.

Scale anchors matter because they set the response frame. In any subjective measure, it remains a challenge to ensure that all respondents understand scale anchors in the same way. It appears to be advisable to adopt absolute scale anchors that encompass the full spectrum of possible experiences (i.e. "never/always/completely" rather than "very", for example) so that there is at least conceptual clarity about where the scales end, even if respondents still define these end states subjectively. Although the difference that this approach is likely to make has not been quantified, it is less ambiguous than the alternatives, and therefore preferable. It may also be advisable to avoid *agree/disagree, true/false* and *yes/no* response formats in the measurement of subjective well-being due to the heightened risk of acquiescence and socially desirable responding – although more concrete evidence on the difference this makes would be welcome.

In terms of labelling the various points along a scale, consistency of approach is essential, and there are evidently benefits and drawbacks to both numerical and verbal labels. Numerical labelling enables a greater number of response options to be used, which may be particularly advantageous when single-item questions are used to assess complex constructs such as life evaluations, without over-burdening respondents (especially via telephone interviewing). Numerical scales are also likely to pose fewer translation problems, which is important to international comparability. For these reasons, in the context of short subjective well-being measures for inclusion in national surveys, numerically-labelled scales are likely to be preferable.

Unipolar versus bipolar measures

The issue

Linked to the issue of scale anchoring is whether the scale is intended to be unipolar (i.e. reflecting a single construct running from low to high) or bipolar (running between two opposing constructs). In a unipolar format, the scale midpoint is intended to represent a moderate amount of the variable of interest, whereas in a bipolar format the midpoint is intended to represent a more neutral territory in between the two opposing constructs:

A unipolar scale:

0	1	2	3	4	5	6	7	8	9	10
Not at all happy					(Moderately happy)					Completely happy

A bipolar scale:

0	1	2	3	4	5	6	7	8	9	10
Completely unhappy					(Neither happy nor unhappy)					Completely happy

Although this distinction between bipolar and unipolar scales may seem very subtle, it should, in theory, have significant consequences for the meaning of the scale points. For example, a score of 0 on the unipolar scale above implies the absence of happiness, whereas a score of 5 implies a moderate amount of happiness. Conversely, on the bipolar scale, a score of 0 should denote complete *un* happiness, a score of 5 implies the respondent is neither happy nor unhappy, and a score around 7 or 8 would imply a moderate amount of happiness. If scale polarity – and the meaning of the midpoint value – is not completely clear to respondents, they may vary in how they interpret the scale, thus introducing a source of error. Furthermore, if one study adopts a 0-10 bipolar scale set of anchors, and the other a 0-10 unipolar set, mean values on these measures may not be comparable, despite both adopting 11-point response scales.

The evidence

Evidence suggests that respondents may have difficulty understanding the intended polarity of affect scales. Russell and Carroll (1999) and Schimmack, Böckenholt and Reisenzein (2002) suggest that many affect scales seemingly designed as unipolar measures could in fact be ambiguous or interpreted by at least some respondents as bipolar. For example, Russell and Carroll (1998; reported in Russell and Carroll, 1999) ran a pre-test with 20 respondents drawn from the general public using the question below, asking them to supply a word to describe each response option:

Please describe your mood right now:

1	2	3	4	5	6	7
Not happy						*Happy*

None of their sample interpreted this as a unipolar scale. The typical respondent placed the scale neutral point (i.e. the absence of happiness) around the middle of the scale (response option 4). All respondents used negative words (e.g. *sad, glum, bad*) to describe response option 2 – thus implying they perceived the response format to be bipolar.

Some response formats may be more confusing than others in terms of communicating scale polarity to respondents. In a study with $N = 259$ undergraduate students, Schimmack et al. (2002) presented a range of different response formats for measures of current affect. All of the measures were intended to be unipolar – that is to say, they were designed to measure one aspect of affect only (*joyful, depressed, pleasant, unpleasant, cheerful, downhearted*, etc.). The majority of respondents, however, indicated that the neutral absence of emotion was represented by the middle of the scale for all four response formats tested. Even with the simplest format of all (the *yes/no* option), only 9% of respondents thought that the absence of emotion was best represented by the *no* category. The measure that divided respondent opinion most of all was the 7-point intensity scale, where 59% of respondents indicated that the absence of emotion was best represented by the scale midpoint (3, labelled *moderately*), whereas only 27% indicated that it was represented by the endpoint (0, labelled *not at all*).

Segura and González-Romá (2003) used item response theory to examine how respondents interpret response formats for affect experienced in the past two weeks. They tested a series of positive and negative affect items with four independent samples, and included both a 6-point frequency (*never – all of the time*) and a 7-point intensity (*not at all*

– *entirely*) response format. Although these response formats are typically used as unipolar measures, Segura and González-Romá's results indicated that respondents tended to construe the formats as bipolar, regardless of whether they required frequency or intensity judgments. One possible interpretation offered for these results is that respondents might use a mental representation of affect that is bipolar.

Although there is a considerable literature regarding the uni- versus bi-polar nature of affect measures, there has been less discussion about polarity or dimensionality in relation to evaluative and eudaimonic measures of subjective well-being. Current practice in single-item evaluative life satisfaction and Cantril Ladder measures is to adopt bipolar extremes as anchors (i.e. "completely dissatisfied/satisfied" and "worst possible/best possible life"); similarly, Diener and Biswas-Diener (2009) and Huppert and So (2009) both anchor their eudaimonia or psychological well-being scales between *strongly agree* and *strongly disagree*.

In a rare study addressing scale polarity in the context of evaluative measures, Davern and Cummins (2006) directly compared unipolar and bipolar measures of life satisfaction and life dissatisfaction. A random sample of 518 Australians completed assessments of life as a whole, as well as seven other sub-domains (e.g. health, personal relationships, safety). The unipolar response format was an 8-point scale, ranging from *not at all satisfied* to *extremely satisfied* (or *not at all dissatisfied* to *extremely dissatisfied*), and the bipolar scale ranged from -7 *extremely dissatisfied* to +7 *extremely satisfied*. The authors reported no significant difference in satisfaction scores derived from the unipolar and bipolar measures, both of which indicated mean scores of around 70% of scale maximum, across the majority of questions. This suggests that respondents essentially treated unipolar and bipolar satisfaction measures the same way. UK experience suggests similar results when comparing a single-item life satisfaction question between surveys conducted by a UK government department (DEFRA) using a bipolar format, and surveys conducted by the ONS themselves using a unipolar format (DEFRA, 2011).

In contrast, the *dis*satisfaction measures reported by Davern and Cummins did differ substantially between response formats. On the unipolar dissatisfaction scale, respondents reported dissatisfaction at around 30% of scale maximum – approximating the reciprocal mean of life satisfaction scores, and thus suggesting dissatisfaction is the mirror opposite of satisfaction (unipolar responses also indicated a strong negative correlation between satisfaction and dissatisfaction). The bipolar scale, on the other hand, mean dissatisfaction was around 65% of the scale maximum, which is difficult to interpret in the light of satisfaction results. The authors speculate that this difficulty with the bipolar dissatisfaction measure may arise from a positivity bias, which focuses respondent attention on the positive anchor of the bipolar measure. The results imply that bipolar scales may cause more problems for negative constructs. This requires further investigation.

Findings across both affective and evaluative measures suggest that respondents do not necessarily fully attend to or fully process scale anchors. On the one hand, this could imply that scale polarity doesn't actually matter too much: even if a scale is constructed with a unipolar response format, respondents might tend to treat it as bipolar anyway. On the other hand, it introduces the obvious risk that with unipolar measures in particular, respondents may differ in how they interpret the measure (e.g. Schimmack et al., 2002). Meanwhile the work of Davern and Cummins indicates that the bipolar *dis*satisfaction scale might have confused respondents.

To reduce difficulties in scale interpretation, the polarity of the response format should be as clear as possible. One clue to scale polarity is the scale numbering adopted. Schwarz et al. (1991) found that an 11-point scale labelled -5 to +5 is more likely to be interpreted by respondents as being bipolar than one using positive numbers only. However, the work of Schimmack et al. and of Russell and Carroll suggests that the opposite is not true. Use of positive numbers alone (e.g. 0 to 11) is not sufficient to cue unipolarity in affect measures. Meanwhile, the bipolar dissatisfaction scale that appeared to confuse Davern and Cummins' respondents already included -7 to +7 scale labels.

It has been suggested that one way to capture unipolar affective constructs could be to begin with a simple *yes/no* question about whether an emotion has been experienced. Both Russell and Carroll (1999) and Schimmack et al. (2002; Study 2) argue that the best way to convey unipolarity in affect measures is first to ask respondents whether or not they feel a particular emotion using a *yes/no* response format, and then ask them to rate the intensity of that emotion, if reported, on a numerical scale with the midpoint clearly labelled. This introduces the possible risk of information loss – in the sense that individuals who respond *no* on a binary measure may still have reported some very slight feelings on an intensity scale – and this requires investigation. Other risks associated with these 2-step "branching" questions are further discussed in the section that follows on the order and presentation of response categories (e.g. Pudney, 2010). One further alternative for shorter and less in-depth measures is to provide respondents with a list of emotions and ask a simple *yes/no* question about whether respondents experienced a lot of those emotions on the previous day. This approach has been adopted in the Gallup World Poll, although the assertion of Green, Goldman and Salovey (1993) that this could increase acquiescence should also be investigated. Dolnicar and Grün (2009) meanwhile found that *yes/no* response formats are not subject to the same patterns of extreme and moderate responding that longer numerical scales can exhibit – thus, the risk of acquiescence might need to be traded off the risk of other forms of response bias.

Key messages on unipolar and bipolar scales

The literature on scale polarity is relatively sparse and largely limited to affect measures, rather than evaluative or eudaimonic measures of well-being. Although not widely studied, there is some evidence to support the view that scale polarity matters, in particular for affective measures of well-being. As one goal of affect measurement may be to test the determinants of different affective states, unipolar measurement scales are often preferred. Given that the evidence implies a general default tendency to interpret affect measures as bipolar, steps may be required to convey this unipolarity more strongly. Options include asking respondents to make simple yes/no judgments about a range of affective experiences, although the risks to scale sensitivity and acquiescence need to be considered.

Many existing and widely-used scales for both life evaluations and eudaimonia are bipolar. Given the apparent tendency of respondents to interpret these types of measures in a bipolar manner, regardless of the actual question wording, adopting bipolar scale anchors may be the least ambiguous in terms of all respondents interpreting the scale in the same way. This suggests that existing bipolar scale structures are probably adequate for these classes of measure, although Davern and Cummins' work does suggest that bipolar response formats may cause problems when explicitly attempting to measure negative constructs (i.e. *dissatisfaction*). There are some grounds to think that adopting negative and positive numerical labels for scale intervals (e.g. -5 to +5) could further

reinforce the bipolarity of measures, but as the additional advantage of this approach is likely to be small, there seems less rationale to depart from current practice – which would in itself reduce comparability with previous work.

Order and presentation of response categories

The issue

The order of response categories may affect which category is selected by default if respondents are satisficing rather than optimising their answers. This could have an impact on the comparability of scores if the order in which response categories are presented varies between surveys. The impact that the ordering of responses has may also vary according to survey mode, thus affecting the inter-mode comparability of the data collected.

The presentation of response categories, and particularly the practice of splitting more complex questions into two distinct steps to simplify them for verbal and telephone interviews, may have also an impact on the comparability of results obtained via different methods.

The evidence

According to Krosnick (1991, 1999), when response alternatives are presented visually, such as on a self-administered questionnaire, satisficing respondents[9] can have a tendency to select earlier response alternatives in a list (sometimes referred to as a *primacy effect*). Krosnick suggests that this is due to a confirmatory bias that leads respondents to seek information that supports response alternatives, and to the fact that after detailed consideration of one or two alternatives, fatigue can set in quite rapidly. This fatigue could in turn then lead respondents to satisfice and opt for the first response category that seems reasonable rather than carefully considering all the possible response alternatives.

By contrast, when response alternatives are read aloud by an interviewer, *recency effects* (where respondents have a tendency to select later response alternatives in a list) are thought to be more likely. This is because the earliest-presented response options can fade out of working memory (or get replaced by new information), and as such they are no longer accessible to respondents.

The key message in both cases seems to be that only a limited number of response categories should be presented to respondents if primacy and recency effects are to be avoided. Where these limits lie is discussed in the section above concerning the number of response options to offer. In addition, study mode can influence the expected direction of effects – and thus, for lengthy questions with a relatively large number of response options, data may be distributed differently among respondents interviewed in different modes. Krosnick (1999) also cites a number of studies indicating that response category order effects are stronger among respondents with lower cognitive skills, and that order effects become stronger as a function of both item difficulty and respondent fatigue.

Bradburn et al. (2004) also argue that if more socially desirable response options are presented first in a list – particularly in the physical presence of an interviewer – respondents may select one of these by default rather than attending to the full list of response choices.

In converting long or complex visual scales for use in telephone and face-to-face interviewing, measures are sometimes divided into two steps, with the question branching into different response categories, depending on how the first step is answered. A practical

example of a branching question, drawn from telephone interviews described in Pudney (2010), is as follows:

Step i): How dissatisfied or satisfied are you with your life overall?

Would you say you are: (1. Dissatisfied; 2. Neither dissatisfied nor satisfied; 3. Satisfied).

Step ii): [if dissatisfied or satisfied…]

Are you Somewhat, Mostly or Completely [satisfied/dissatisfied] with your life overall? (1. Somewhat; 2. Mostly; 3. Completely).

Pudney (2010) provides evidence to suggest that 2-step branching questions may significantly alter response distributions to satisfaction questions. In survey data examining overall life satisfaction, job satisfaction, and satisfaction in health, household income and leisure time, response distributions among women were significantly different for every domain except income when the 2-step branching procedure was used. Among men, the branching format only had a significant impact on the distribution of responses on the job satisfaction measure. In general, the 2-step branching questions tended to result in higher overall satisfaction assessments – with a higher frequency of extreme values selected. There were also some significant differences in the relationships between life circumstances and the health, income and leisure satisfaction scores when these outcomes were assessed using a 2-step question structure, as compared to the 1-step structure. In particular, the coefficient for household income, which was large and significantly different from zero when income satisfaction was measured using the 2-step procedure, became very small and insignificant when income satisfaction was measured using a 1-step approach.

While Pudney found that responses differed between 1-step or 2-step questions, it is not clear from this research which question structure is *best* in terms of either the accuracy or reliability of the measure. As noted earlier, it has been hypothesised that a 2-step question structure may actually make it easier to measure positive and negative aspects of affect independently from one another (Russell and Carroll, 1999; Schimmack et al., 2002), which is of theoretical interest, even if it is longer and more cumbersome for both respondents and interviewers. These trade-offs need to be better understood.

A further issue for the presentation of response categories is where a battery of several questions requires some mental switching on the part of respondents between positive and negative normative outcomes. For example, if a 0-10 response scale, anchored with 0 = *not at all* and 10 = *completely*, is used to assess *happiness yesterday*, a high score represents a normatively "good" outcome, and a low score represents a normatively "bad" one. If the same response format is then used immediately afterwards to measure *anxiety yesterday*, a high score represents a normatively "bad" outcome, and a low score represents a normatively "good" one. Rapid serial presentation of such items risks causing some degree of confusion for respondents, particularly in the absence of visual aids or showcards.

One initial piece of evidence from cognitive testing has suggested that respondents can sometimes struggle to make the mental switch between questions that are framed positively and negatively. The ONS (2011c) looked at this issue using the 0-10 *happiness yesterday* and *anxiety yesterday* response format described above. In the experimental subjective well-being question module tested by the ONS, the two affect questions are preceded by two positively-framed questions about life evaluations and eudaimonia, making anxiety yesterday the only question where 0 is a normatively "good" outcome. Their findings indicated that some respondents treated the scale as if 0 = "bad outcome",

10 = "good outcome", regardless of the item in question. Further research is needed to see whether alternative response formats, question ordering or a greater balance between positively- and negatively- framed items can alleviate this difficulty without over-burdening respondents. The impact of question order more generally will be examined in detail in Section 3.

Key messages on the order and presentation of response categories

While there is evidence that scale order can impact results when verbally-labelled scales are used, the reliance on numerical rather than fully-labelled scales for many measures of life evaluation and affect implies that scale ordering is unlikely to be a significant problem for these measures. There is, however, merit in presenting even numerical response order consistently, such that scales run from 0-10 (the classic presentation), rather than 10-0.

Complex questions are sometimes broken down into two separate steps for the purposes of telephone presentation. There is some evidence to suggest that this can alter the overall pattern of responses. As the vast majority of established subjective well-being measures adopt a normal 1-step phrasing structure, it seems preferable to maintain this wherever possible for the sake of comparability (and to reduce survey complexity, both for respondents and interviewers). If compelling evidence emerges to suggest that the 2-step procedure is preferable in some circumstances – and this is perhaps most likely in the case of short-term affective subjective well-being measures – then their use could be further reconsidered. If response categories are limited to simple numerical scales (as recommended for life evaluations as a minimum), these should be less challenging to deliver in telephone interviews, as respondents are required to remember fewer verbal labels. Thus breaking these questions down into two parts may be unnecessary.

The impact of asking respondents to perform mental switching between the underlying "good" and "bad" normative direction of response formats needs further research. The relative merits of frequency (*never… all the time*), intensity (*not at all… completely*), and binary (*yes… no*) response formats for affect items also needs to be investigated systematically, with and without verbal and numerical labels for the scale points, and with reference to the impact on scale sensitivity. Other issues with regard to question ordering, including whether to ask positive or negative items first in a battery of questions, are discussed in the section on question order that follows.

Cross-cutting issues and overall messages on response formats

The variety of different response format options available lead to a wide array of possible combinations, and it is thus worth drawing conclusions together. One of the difficulties in interpreting the evidence in this field is that few studies have taken a systematic approach to how these combinations are tested – thus, investigations of scale length may fail to test whether scale polarity matters for optimal scale length. Meanwhile, examination of whether to add verbal labels to a scale might find that verbal labels increase measurement reliability – but gives no indication as to whether this applies equally to both 5-point and 11-point scales, or to both frequency and intensity judgements. Similarly, survey mode is often neglected in this field, so that it cannot be known with certainty whether conclusions drawn on the basis of pen-and-paper surveys can be transferred to face-to-face or telephone interviews, and vice versa.

There are also trade-offs that need to be considered. For example, verbal labels might increase test-retest reliability (perhaps by making differences between response categories more salient for respondents), but verbal labels may in themselves be a source of confusion and/or variability in how scales are interpreted, both between different individuals and between different languages.

Nevertheless, decisions must be taken on the basis of existing evidence, not on an (unavailable) ideal evidence base. Furthermore, it is not clear that the existing evidence base regarding the optimal question structure for measures of subjective well-being is any worse than for many other topics commonly collected in official statistics.

Considering the available evidence on response formats, several conclusions emerge:

- Response format does matter. Use of different response formats can introduce non-trivial variance in comparisons between measures intended to capture the same underlying concept. There is therefore a strong *prima facie* case for consistency in measurement. This is particularly important for key national measures that are likely to form the basis for international comparisons.
- There is clear evidence in favour of longer (7 to 11 point) scales over shorter (2 to 5 point) scales for single-item measures of life evaluation, and several recent high-quality studies suggest that an 11-point scale has significant advantages in terms of data quality. This lends significant weight to the use of the 11-point 0-10 scale already used in a number of prominent unofficial and official surveys. Evidence regarding optimal scale length for affective and eudaimonic measures is lacking. Another important question is which formats respondents tend to prefer.
- In selecting scale anchors, there is a preference for verbal labels that denote the most extreme response possible (e.g. *always/never*). Concerns regarding the use of *agree/disagree, true/false* and *yes/no* scale anchors in relation to response biases such as acquiescence and social desirability should be further tested.
- Linked to the issue of scale anchors is the question of whether response formats should be constructed as unipolar (*not at all – completely*) or bipolar (*completely dissatisfied – completely satisfied*). The evidence available indicates that a sizeable proportion of respondents may have a tendency to treat unipolar measures as if they were bipolar. In the case of affect, separate measures of positive and negative affect are often desirable, and there extra effort may be needed to convey the unipolarity of scales.
- For life evaluations and eudaimonia, the majority of existing measures are bipolar scales. There is limited evidence in this area, but what evidence is available suggests that whilst the choice between unipolar and bipolar appears to make little difference to positively-framed questions (such as satisfaction), bipolar formats for negatively-framed questions may prove more problematic.
- Providing only numerical labels for scale intervals is likely to allow simpler transferability between languages, which is an important consideration for international comparability. Verbal labels can meanwhile help to convey meaning – but generating them could be very challenging for subjective measures with up to 11 response categories (and these longer scales may be desirable for single-item measures in particular). Providing a numerical (rather than verbal) label for all scale intervals is therefore advised.

- In terms of the order and presentation of response categories, it is important to keep measures simple and to facilitate the comparability of data across survey modes. If numerical rather than verbal labels for response categories are adopted on a sliding scale, the order of presentation to respondents is not expected to be a significant issue, although consistency in running from 0-10, rather than 10-0, is recommended. Due to their additional complexity, 2-step branching questions are not recommended for subjective well-being measures at present, although if further research demonstrates that they have particular advantages – for example, in the separate measurement of positive and negative affect – this could be reconsidered.

One important practical question is the extent to which it is necessary or desirable to present respondents with the *same* response format across a battery of subjective well-being questions, i.e. in a module designed to assess life evaluations, affect and eudaimonia, is it necessary to use a consistent response format? If one of the analytical goals is to compare responses on two questions directly (for example, to test the effect of question wording), it is likely to be important that those questions use an identical response format (particularly in terms of the number of response options offered). For example, there may be some advantages in being able to directly compare single-item life satisfaction and eudaimonia questions, given their complementary nature.

Having an identical response format may be less important when the goal is to compare single-item life evaluations with multiple-item affect and eudaimonia measures. The impact of a change in response formats, however, also needs to be considered. On the one hand, a shift in the response format could act as a cue to respondents that a new topic is now being examined, thus reducing the risk of shared method variance (i.e. in this case, respondents reporting in a similar way on two separate measures simply as a result of those measures sharing the same response format). This could also assist in enabling respondents to make the mental switch between a set of life evaluation and eudaimonia questions that tend to be positively-framed, and a set of affect questions that contain a mix of positively- and negatively-framed items. On the other hand, changing the response format will require an explanation of the new question and response categories, which takes up survey time and effort on the part of the respondent. As with any survey item, there may be a trade-off between the "ideal" question structure and analytical, presentational and practical convenience. Further evidence from large-scale investigations can help to indicate where the most sensible balance lies.

3. Question context, placement and order effects

Introduction

The survey context, such as question order, introductory text and the survey source, can influence respondents' understanding of individual questions within a survey, as well as the information that they draw on in order to answer those questions. Often, this sensitivity to context can serve to enhance respondent understanding, and thus improve the quality of data obtained. However, as aspects of the survey context can also have unforeseen and undesirable consequences – introducing bias and affecting the comparability of data collected in different contexts – an agreed approach on survey context is also important for data quality, validity and comparability.

The main concern is that features of the survey context might inadvertently cause respondents to misinterpret question meaning or bias the manner in which responses are constructed – including by making certain types of information more accessible to respondents than others. For example, a set of questions about difficult life events (such as recent unemployment or bereavement), if asked immediately prior to subjective well-being questions, could set the affective "tone" for the questions that follow and/or signal to respondents that they should take this information into account when forming their answers.

The process of answering a survey may also trigger either conscious or sub-conscious self-presentational behaviour, such as the drive to appear consistent across responses, or social desirability effects (i.e. a tendency to present oneself in a favourable light, and/or give responses that conform to prevailing social norms). Experimental demand effects, whereby respondents act in accordance with their own implicit theories, or attempt to second-guess the surveyor's views about how survey items may be related to one another, can also influence patterns of responding across groups of items and interact with question order effects. For example, a survey about health-related disabilities could cause respondents to focus on health-related information when answering more general subjective well-being items, because this is what they perceive to be expected of them.

It has been argued that the nature of some subjective well-being questions can mean that, rather than retrieving a specific piece of information from memory, survey respondents are likely to construct their answers on the spot, making them particularly susceptible to context effects (Sudman, Bradburn and Schwarz, 1996; Schwartz and Strack, 2003). According to this logic, evaluative and eudaimonic questions are perhaps at greatest risk of context effects, given that they are more all-encompassing in their wording, and respondents may therefore seek contextual information to help them understand exactly what sort of response is required. Affect measures meanwhile refer to something more tangible and directly rooted in the recent past, thus respondents may have stronger representations in memory to draw on.

Question context and the impact of question order

The issue

A key concern often raised in the literature is that preceding questions may affect how respondents interpret the meaning of an item and/or the type of information that is temporarily accessible to respondents when constructing their answers – effects often collectively referred to as *priming*. The precise influence of priming effects is not always simple to predict or easy to detect, however, and there may be individual differences in the extent to which question context exerts an influence. For example, asking about marital status immediately before a life satisfaction question may not exert a strong influence on respondents whose marital status has remained stable for a number of years, but may be more salient for those recently married, divorced or widowed – potentially evoking positive feelings in some individuals, but more negative ones in others.

Context effects may therefore exert an influence on both the mean level of a measure, when summed across respondents, and/or the distribution of data, when context effects impact differently among different subgroups within a sample. This can in turn threaten both the comparability of data from different surveys with differing contexts, as well as the comparability of data from different groups of respondents on the same survey. Context effects may also inflate or suppress relationships between variables by making certain

information and/or mood states more accessible to respondents. For example, if a set of prior questions about health status prompts respondents to draw more heavily on health-related information when answering a subsequent life satisfaction question, this could lead to a higher correlation between health status and life satisfaction than one might find in other surveys where the questions are arranged differently.

The evidence

Priming effects are thought to occur when preceding questions (or other contextual cues[10]) make certain information or emotional states more accessible to respondents, influencing how they answer subsequent questions. The expected direction of influence, however, can vary under different circumstances. *Assimilation effects* refer to priming where subsequent responses are consistent with the information or emotions that have been made more accessible by contextual factors (for example, if prior questions about recent positive life events prime an individual to respond more positively to a life evaluation question). *Contrast effects* meanwhile refer to priming effects where subsequent responses contrast with the information or emotions made temporarily accessible by contextual factors. Sometimes these contrast effects are thought to occur because the prime serves as a comparison reference-point for respondents. For example, recalling distant positive events before reporting current life satisfaction might induce more negative perceptions of life currently, if the comparison is not favourable (Schwartz, 1999).

Across several studies, Schwartz and colleagues have reported that respondents' answers to subjective self-report questions can be sensitive to adjacent questions in a survey (Schwartz, 1999; Schwartz and Strack, 2003). Whilst in telephone and face-to-face interviews priming is restricted to the effect of preceding questions on subsequent questions, in the case of self-administered questionnaires (where respondents can flip back and forth between questions), priming can also work in the opposite direction (Schwartz, 1999).

One source of concern for survey design is that asking about life events might prime respondents to place undue emphasis on these events when subsequently reporting life evaluations or experienced affect. In a classic but small-scale ($N = 36$) priming study, Strack, Schwarz and Gschneidinger (1985) examined the circumstances under which describing life events might influence subjective well-being data. They found that respondents who were asked to provide vivid and detailed information about three recent positive events subsequently reported higher positive affect and higher evaluative subjective well-being than respondents asked to provide vivid and detailed information about three recent negative events. Respondents instructed to provide less vivid and detailed accounts did not show this pattern of results.

Some research has also suggested that priming can influence the kinds of information that respondents rely on when forming evaluative subjective well-being judgements. For example, Oishi, Schimmack and Colcombe (2003) systematically primed "peace" and "excitement" by asking a small group of student respondents to read a fictional obituary and rate either how peaceful or exciting the deceased's life was. Respondents then completed the 5-item Satisfaction with Life Scale, and rated the extent to which they had experienced 16 different emotions in the past month – including excitement. Oishi and colleagues found that the priming condition changed the basis on which life satisfaction judgements were made – such that excitement scores were more strongly related to overall

life satisfaction in the excitement-primed condition [$r(37) = 0.63$, $p < 0.01$] than in the peaceful-primed condition [$r(38) = 0.29$, *ns*]. There were, however, no significant mean differences in life satisfaction between the excitement- and peaceful-primed conditions.

There is also some limited evidence suggesting that sensitivity to context effects may vary across situations and cultures. For example, Haberstroh et al. (2002) hypothesise that circumstances or cultures that emphasise interdependence with others (such as the Chinese culture) may cause people to pay more attention to the *common ground* between the questioner and the respondent, thus creating greater susceptibility to context effects. This was tested and supported in studies by Haberstroh et al. on the impact of question order on subjective well-being (e.g. paradigms from Strack, Schwartz and Wänke, 1991; and Schwartz, Strack and Mai, 1991). If robustly replicated, this suggests differential sensitivity to question order effects could influence the comparability of international findings.

Asking socially sensitive questions immediately prior to evaluative judgements can also influence how respondents form their answers. Within a large national survey context, Deaton (2011) found evidence of strong question order effects in the Gallup Healthways Well-being Index data from the United States during 2008. Specifically, those asked political questions before evaluative subjective well-being measures (using the Cantril Ladder, on a 0-10 scale) on average reported well-being around 0.6 of a rung lower than those not asked any political questions. This effect is only a little smaller than the effect of becoming unemployed, and about the same size as the decline in Ladder scores recorded over the course of 2008 – a period during which the world financial system plunged into crisis. It was not replicated, however, in the daily affect measures taken in the same survey. Respondents were only very slightly more likely to express negative emotions when political questions were included. Positive affect measures were totally unaffected by the inclusion of political questions. This could imply that affect measures are more robust to the impact of (political) question context than evaluative measures, although it is also important to note that the affect measures were positioned much later in the survey than the evaluative measure (which was immediately adjacent to the political questions right at the beginning of the survey).

The buffering impact of intervening text or questions can help to reduce context effects. Deaton (2011) reported that the impact of political questions on life evaluations dropped markedly when Gallup added a transition or "buffer" question between political questions and life evaluations.[11] Later data showed that removing the political question about *satisfaction with the way things are going in the US* (and retaining only one political question – the *Presidential approval rating*) produced indistinguishable scores between those respondents who were asked political questions and those who were not.

The inclusion of "buffer items" between life evaluations and life events also appears to reduce the likelihood of any "contamination" between them. In a laboratory-based study, Sgroi, Proto, Oswald and Dobson (2010) measured overall happiness (on a 0-7 scale) at the beginning of the study and then gave participants a series of distraction tasks. Participants were then asked priming questions about life events, before completing several buffer items, followed by a life satisfaction question (on a 0-10 scale). Whilst illness and positive events were very significantly associated with overall life satisfaction, they were also very strong predictors of the initial happiness question (which could not have been contaminated). Crucially, with initial happiness levels controlled, the relationship between events and satisfaction became non-significant – indicating that the magnitude of the

event-satisfaction relationship was genuine and linked to the overall influence of events on global subjective well-being, rather than being artificially inflated due to the inclusion of primes. Sgroi et al. therefore reported that subjective well-being measures are robust to priming effects from important life events.

The operation of transition questions or buffer items critically depends on the *nature* of the buffers used. Bishop (1987) found that being unable to answer two political knowledge questions decreased self-reported interest in public affairs, and this effect persisted despite the inclusion of 101 unrelated intervening items. However, following Bishop's study design, Schwarz and Schuman (1997) showed that a single buffer item that provided respondents with an alternative (external) explanation for their lack of knowledge greatly reduced the context effects. Schwarz and Schuman concluded that when buffer items are unrelated to the question of interest, the only way they can reduce context effects is by attenuating the accessibility of information brought to mind by the context effect. Related buffer items, on the other hand, can change the *perceived implications* of the context information – and although this sometimes reduces context effects, in other circumstances it could *increase* the impact of context information by making it "more diagnostic".

Further work is needed on the most effective forms of buffers for subjective well-being questions. On the one hand, Bishop's work implies that where context information is highly salient, interference from even a large set of unrelated buffer items will not necessarily reduce that salience. On the other hand, Schwarz and Schuman (1997) suggest that buffers *related* to subjective well-being questions could cue certain types of information and produce context effects of their own, further complicating the picture.

Key messages on question context and the impact of question order

Available evidence suggests that question context – and particularly question order – can influence subjective well-being reports, and that in some cases these effects can be large. Given that subjective well-being questions may be included in a range of different surveys covering different subject matters, identifying and applying the best approach for minimising survey context effects should be a priority. Producing measures that are robust to context effects is important for ensuring data comparability – across surveys and across countries.

Locating subjective well-being questions as early on in the survey as possible should limit interference from other items. However, the novel nature of subjective well-being questions may come as a shock to respondents if presented right at the beginning of a household survey, before interviewers have had the opportunity to build some degree of rapport. It is most important to avoid placing subjective well-being questions immediately after questions that are likely to elicit a strong emotional response. Use of introductory text and transition questions may also help to reduce context effects. These and other practical recommendations are discussed further in Chapter 3.

Finally, it is unlikely that subjective well-being measures are uniquely susceptible to context effects, and the impact of prior subjective well-being questions on responses to subsequent questions on *other topics* (such as self-reported health status, or subjective poverty) remains an important area for further study.[12] Until more is known about this impact, it seems strongly advisable to maintain some distance between these items in the survey through the use of more neutral transition questions and introductory text that clearly distinguishes between question topics.

Question order within a subjective well-being module

The issue

Order effects may also be an important consideration for question order *within* a module or cluster of subjective well-being questions. For example, asking a set of positive affect questions might prime positive emotions that could influence subsequent answers to life evaluation, eudaimonia or negative affect questions. Similarly, a drive for consistency may mean that respondents who report positive life evaluations overall might then feel that subsequent eudaimonia questions should also be answered positively.

Order effects can also have implications for the overall *number* of subjective well-being questions that should be included. Strack, Schwartz and Wänke (1991) have discussed the potential importance of the conversational principle of non-redundancy – i.e. that partners in a normal conversation will tend to avoid asking the same question, or providing the same information, more than once. Thus, if someone asks a question similar to one asked only moments earlier, respondents might assume that different information is required, creating contrast effects. This could lead to correlations among a set of subjective well-being questions being artificially suppressed if respondents assume that each question must require a different response.

The evidence

One of the most oft-cited and oft-studied of all context effects relates to the impact of asking about personal relationships, such as dating frequency or marital happiness, prior to asking evaluative subjective well-being questions (e.g. Strack, Martin and Schwarz, 1988; Schwarz, Strack and Mai, 1991; Schuman and Presser, 1981; Smith 1982, cited in Tourangeau, Rasinski and Bradburn, 1991). Some of these studies find correlational effects – i.e. stronger relationships between personal relationships and overall life satisfaction when the question about personal relationships is asked first – but not directional effects, i.e. no differences in mean scores or the percentage of *very happy* respondents (e.g. Tourangeau, Rasinkski and Bradburn, 1991; Schwarz, Strack and Mai, 1991). Other studies have however found evidence of mean score differences, such that answering questions about marriage satisfaction induces higher happiness ratings overall, but produces no change in the correlation between marital and life satisfaction (Schuman and Presser, 1981; Smith 1982).

Variability across results, both in terms of the direction of effects and their magnitude, appears to be quite persistent. For example, following the procedure used by Strack, Martin and Schwarz (1988), Pavot and Diener (1993a) found much weaker context effects among single-item life evaluation measures, and no effect of context on the multi-item Satisfaction with Life Scale. Schimmack and Oishi (2005) performed a meta-analysis of all known studies exploring the effect of item order on the relationship between an overall life satisfaction measure and a domain-specific satisfaction measure. Sixteen comparisons from eight different articles were examined. Overall, the meta-analysis indicated that item-order effects were statistically significant ($z = 1.89$, $p < 0.02$), but the average effect size was in the "weak to moderate range" ($d = 0.29$, $r = 0.15$). Like Tourangeau et al., the authors also noted that the results were extremely variable across studies, with effect sizes ranging from $d = 1.83$ to -0.066. Further empirical investigation by Schimmack and Oishi reaffirmed that overall item-order effects were small or non-significant, but also that it is difficult to make *a priori* predictions about when item-order effects will emerge. However,

they did find that priming an irrelevant or unimportant domain (such as weather) produced no item order effects, and similarly, priming a highly important and chronically accessible domain[13] (such as family) also failed to produce item order effects.

The implication of Schimmack and Oishi's findings is that item order effects should be most likely when preceding questions concern domains of life that are *relevant* to an individual's overall life satisfaction, but that are chronically accessible to the individual in only a weak way. For example, satisfaction with recreation might be *relevant* to overall life satisfaction, but might not be something that always springs to mind when individuals make life satisfaction judgments. Asking a question about recreation prior to one about overall life satisfaction might make recreation-related information more accessible and more salient, thus strengthening the relationship between this and overall life satisfaction. Schimmack and Oishi (2005) tested this hypothesis in relation to housing satisfaction, but failed to show significant order effects. However, this may be because of large individual differences in how important, relevant, and chronically accessible housing information is. For example, Schimmack et al. (2002) found that housing was highly relevant for some individuals and irrelevant for others.

Tourangeau et al. also speculate that some of the variability among findings may be due to the introduction given to the questions that immediately precede the satisfaction questions (e.g. questions about marital status) and to whether marital happiness/satisfaction is one of many other domains assessed alongside overall life satisfaction (because the effect may be reduced if there are several domain-specific items preceding the overall judgment).

Although the picture provided by this work is a complicated one, the available evidence on item-order effects suggests that, to ensure some consistency of results, general life evaluation questions should precede questions about specific life domains, particularly when only a small number of domains are considered. Furthermore, if demographic questions (such as marital status) are asked before evaluative subjective well-being questions, there should be some introductory text to act as a buffer. Specific instructions to respondents to include or exclude certain domains from overall life evaluations (e.g. "aside from your marriage" or "including your marriage"), however, are not recommended, as these can also influence the pattern of responding in artificial ways (because overall evaluations of life would be expected to incorporate information such as marital satisfaction).

Although most of the work on question order has focused on evaluative subjective well-being judgements, the UK Office of National Statistics (ONS, 2011b) have reported an effect of question order on multiple-item positive and negative affect questions. In a split-sample randomised trial using national data ($N = 1\,000$), the ONS found that asking negative affect questions first produced lower scores on positive affect items – and this effect was significant (at the $p < 0.05$ level) in the case of using adjectives such as *relaxed, calm, excited* and *energised*. Conversely, when positive affect questions were asked first, the mean ratings for negative affect questions were generally *higher* – except in the case of *pain* – and this increase was statistically significant for the adjectives *worried* and *bored*.[14]

On the issue of *how many* subjective well-being questions to ask within a survey module, Strack, Schwartz and Wänke (1991) found that asking questions about two closely related constructs could produce distortions in the data. These authors examined the correlations between evaluative *life satisfaction* and *happiness* questions administered: i) in two separate and seemingly unrelated questionnaires; and ii) concurrently in the same

questionnaire, with a joint lead-in that read, "Now, we have two questions about your life". The correlation between the measures dropped significantly from $r = 0.96$ in condition i) to $r = 0.75$ in condition ii). Strack et al. infer that respondents in condition ii) were more likely to provide different answers to the two questions because they were applying the conversational principle of non-redundancy. Specifically, respondents may assume that two similar questions asked on the same survey must require different responses because asking the same question twice would be redundant.

Key messages on question order within a subjective well-being question module

Although overall effect sizes appear to be small in most cases, the presence of order effects *within* groups of subjective well-being questions has some clear implications for survey design. First, it seems advisable to ask the most general evaluative questions first, followed by domain-specific evaluative questions as necessary. If evaluative subjective well-being is measured by single-item scales, using only one of these measures should reduce redundancy and any potential for respondent confusion or fatigue. This means that a choice must be made between, for example, the Cantril Ladder, a life satisfaction question and an overall happiness question, rather than including them all in one survey. Where domain-specific measures are to be included, covering a wide range of domains should reduce the likelihood of respondents focusing on any one particular domain (such as marital or relationship satisfaction).

The approach of running from the general to the specific suggests that surveys should move from global evaluative measures to eudaimonic questions and then to more specific affect measures – although further work is needed to explore how responses to each of these questions may interact and concerning the buffering text and/or question instructions that might be best used to break up the question module. In the case of affect measures, although in theory randomised presentation of positive and negative affect items is likely to represent the optimal solution, in practice this could heighten the risk of respondent or interviewer error, particularly in less technologically-advanced survey settings. Thus, the best way to ensure comparability of results may be to use a fixed item order across all surveys. This requires further investigation.

Survey source and introductory text

The issue

A final potential source of context effects comes from the survey source itself, and what respondents may – either implicitly or explicitly – infer from this about how they should answer questions. Introductions and framings have an impact on how respondents understand the objectives of the survey. Norenzayan and Schwarz (1999) and Smith, Schwarz, Roberts and Ubel (2006) argue that participants, as co-operative communicators, will try to render their responses relevant to the surveyor's assumed goals and interests (following the conversational principle that information provided to a listener should be *relevant*, Grice, 1975). A less benign interpretation is that survey design can induce experimental demand effects – i.e. where the surveyor's *a priori* hypotheses about the nature of relationships between variables, and/or the respondents' views on those relationships, may influence the pattern of responses. Finally, it is possible that respondents may have reason to present themselves or their experiences in a positive light (socially desirable responding) or use survey responses to communicate a specific message to the surveyor.

If the survey source affects the manner in which respondents answer questions, this could have implications for data comparability, particularly between official and non-official surveys. International differences in how statistical surveys are conducted could also affect comparability. The way a survey or question module is introduced can also play a key part in motivating respondents – and it will therefore be important to do this in a uniform manner.

The evidence

It is difficult to isolate the impact of the survey source on responding. Most information on this effect therefore comes from studies where the survey source is experimentally manipulated. For example, Norenzayan and Schwarz (1999) asked respondents to provide causal attributions about mass murder cases – and they found that respondents were more likely to provide personality- or disposition-based explanations when the questionnaire letterhead was marked *Institute for Personality Research* and more social or situational explanations when it was marked *Institute for Social Research*.

In a further example, Smith et al. (2006, Study 1) found that the correlation between life satisfaction and a subsequent health satisfaction question was higher when the survey was introduced to respondents as being conducted by a university medical centre (focused on the quality of life of Parkinson's disease patients) than when the survey was introduced as being conducted by the university in general (and focused on the quality of life of people in the eastern United States). The health satisfaction component of life satisfaction was much greater in the medical centre condition, accounting for three times as much variation in the life satisfaction measure (39.7% as opposed to 11.5%). Smith et al. liken this to the assimilation effects observed in studies of question order.

In national surveys that cover a very wide range of topics, any *a priori* assumptions that might be held about relationships between variables are likely to be obscured by the sheer breadth of the survey. Short, sharp, opinion-based surveys, on the other hand, might be more likely to be viewed by respondents as "hanging together". So, for example, in the Gallup poll described by Deaton (2011), the questions around the "direction" of the country asked at the very beginning may have set the tone for the subjective well-being questions that followed, and respondents may have been conflating their own subjective well-being directly with their views on national well-being.

Knowing that a survey originates from a national statistical office is unlikely to give respondents many cues as to how they should respond to subjective well-being questions in particular – although knowing that the data is explicitly linked to the information needs of the government may introduce a risk that respondents will tailor their answers (consciously or otherwise) in order to send a message to those in power. It is not clear at present whether or how much of a threat this poses to subjective well-being data. Deaton (2011) noted the close relationship between subjective well-being measures and stock market movements between 2008 and 2010. When national-level data on subjective well-being become available for monitoring purposes, it will be interesting to see whether this follows the political cycle or other potential determinants of national mood. It will also be important to investigate the source of any differences in results between official and unofficial surveys.

Key messages on survey source and introductory text

For ethical reasons, some degree of information about the survey source and its intended purpose must be clearly communicated to respondents. This, however, runs the

risk that some respondents will adapt their answers to provide information they think will be most relevant and/or to communicate a message specifically to the surveyor. There are a number of good reasons why subjective well-being questions should be embedded in a larger national household survey rather than being measured separately – a key one being that it will enable the exploration of relationships between subjective well-being and other policy-relevant co-variates. A further reason is that one might expect reduced effects of survey introduction and source on subjective well-being questions if they are part of a very general and broad-ranging survey, rather than one that specifically focuses on subjective well-being.

The finding that context matters suggests that the same subjective well-being questions included in, say, a national opinion or social survey may elicit different responses than when included in a labour force survey or a survey more heavily focused on objective and economic indicators. Inclusion of a standard set of subjective well-being questions in different national survey vehicles will provide a unique opportunity to test this in the future.

4. Mode effects and survey context

Introduction

This section discusses survey mode and timing as well as the impact of the wider context in which surveys are conducted – such as incidental day-to-day events that might affect responses. It essentially concerns the extent to which subjective well-being data collected under different circumstances using different methods can still be considered comparable. It also examines the question of whether there is an "optimal" method for subjective well-being data collection, in terms of data quality.

Survey mode

The issue

There are a variety of different survey methods available for the collection of subjective well-being data. These include:

- Self-Administered Questionnaires (SAQs), traditionally conducted in a pen-and-paper format, but which increasingly involve Internet-based surveys.
- Computer-Assisted Self-Interviews (CASI), including variants with pre-recorded audio presentation of questions presented through headphones (audio-CASAI).
- Telephone interviews and Computer-Assisted Telephone Interviews (CATI).
- Pen-and-paper interviewing (PAPI) and Computer-Assisted Personal Interviews (CAPI) usually conducted through visits to the survey respondent's home.
- Diary methods, including Time-use Diaries, Experience Sampling and the Day Reconstruction Method (see Box 2.2).
- Computer-Assisted Web-Interviewing (CAWI), although currently this technique cannot be used where nationally-representative samples are required due to strong sampling biases.

The central issue with regard to survey mode is whether data collected in different modes can be considered comparable. In general, survey mode effects can take the form of: 1) *coverage error*, i.e. certain modes excluding or failing to reach certain segments of the population; 2) *non-response bias*, i.e. different respondents having preferences for different modes; and 3) *measurement error* (Jäckle, Roberts and Lynn, 2006). The current chapter is

Box 2.2. **Experience Sampling and the Day Reconstruction method**

Some measures of subjective well-being, particularly affective measures, require respondents to retrospectively recall their previous experiences over a given time frame. A frequent concern is that various self-report biases (including those linked to certain personality traits) can influence this recall process. In terms of minimising the memory burden and the risk of recall biases, *Experience Sampling Methodologies* (ESM – Csikszentmihalyi and Larson, 1992; Hormuth, 1986; Larson and Delespaul, 1992), also known as *Ecological Momentary Assessments* (EMA – Schwartz and Stone, 1998; Smyth and Stone, 2003; Stone et al., 1998) represent the "gold standard". In these methods, respondents provide "real-time" reports throughout the study-period, and the memory burden is either very small (e.g. summing experiences over the past few hours) or nonexistent (e.g. requiring respondents to report how they are feeling *right now*). Studies typically involve between two and twelve recordings per day (Scollon et al., 2003) and may last one or two weeks. To ensure compliance among respondents (for example, to detect and prevent the "hoarding" of responses until the end of the day, which can be a significant problem with paper diaries), it is advisable to use electronic diaries, such as palm-top computers pre-programmed with questions and with an internal clock that can both remind respondents when entries are due and record the timing of responses (Stone et al., 2002).

Whilst experience sampling methods have some significant advantages in data quality, the study design is burdensome for both respondents and research resources. A less intrusive and burdensome alternative is offered by the *Day Reconstruction Method* or DRM (Kahneman et al., 2004). This technique is designed to assist respondents in systematically reconstructing their day in order to minimise recall biases. It builds on evidence suggesting that end-of-day mood reports may be more accurate than previously supposed (Parkinson et al., 1995), and that retrospective accounts of mood may be reasonably valid for periods of up to 24 hours (Stone, 1995). The DRM represents a more pragmatic alternative to the ESM but still requires detailed survey modules, which can for example take respondents between 45 and 75 minutes to complete (Kahneman et al., 2004).

particularly concerned with the third of these issues, and although discussion of sampling issues is covered in Chapter 3, the first and second issues also have consequences that potentially interact with mode effects and problems with data quality – for example, where mode effects are more likely among certain groups of respondents.

A further crucial consideration is the extent to which identical question wording and response formats can be used across different survey modes. As noted in Sections 1 and 2 of this chapter, question wording and response formats can have non-trivial impacts on responses. If questions designed for pen-and-paper questionnaires or face-to-face interviews need to be modified for presentation over the telephone, for example, this may reduce the comparability of the data collected.

Techniques for the measurement of subjective well-being can vary substantially on a wide range of dimensions that may be of relevance to measurement error, such as:

- The extent of human interaction and the extent to which respondents are permitted opportunities to clarify the meaning of questions or response categories.
- Privacy and the risk of audience effects, i.e. whether having other people (such as other household members) present at the time the survey is completed influences how respondents portray themselves.

- The pace with which the survey is conducted and the extent to which the flow of survey questions is determined by the interviewer or by the respondent. This is also connected to the opportunity respondents have to revisit questions or pre-read the questionnaire in advance of completing it.
- The auditory versus visual presentation of questions and response categories and the subsequent memory and information-processing burden placed on respondents.

There are various ways in which the above features of survey mode can influence data quality – for example, through influencing respondent motivation, question comprehension and the likelihood of satisficing and response biases. Interviewer interaction and audience effects are also expected to influence the extent to which self-presentational effects and socially desirable responding are likely to occur, such as presenting oneself in a positive light, or conforming to social norms. In the case of subjective well-being measures, self-presentational biases, if present, would be expected to increase reports of positive evaluations and emotions and decrease reports of negative ones. Although self-presentation effects are quite a wide-reaching issue for data quality, discussion of them is generally limited to this section of the chapter, because the main practical implications in terms of survey methodology concern the impact of survey mode.

Different survey methods have different advantages, and steps taken to reduce one risk to data quality (such as social desirability) may have consequences for other risks (such as other forms of response bias). Furthermore, where mode differences are detected, this does not in itself tell you which mode is producing the more accurate data. Much of the discussion that follows therefore describes the extent to which survey mode appears to influence subjective well-being data and the extent to which data collected in different modes can be compared with confidence.

The evidence

Social desirability. When mode effects are observed on socially sensitive survey items, they are sometimes attributed to social desirability effects. The underlying assumption is that a lack of anonymity, and/or a lack of perceived confidentiality, particularly in interview settings, may cause respondents to report higher levels of socially desirable attributes, including higher subjective well-being. Audience effects, where a respondent gives their answers in the presence of one or more other individuals (aside from the interviewer), can also have a variety of impacts, depending on the nature of the relationship with the audience and the impression a respondent may be seeking to create.

Different views exist regarding the likelihood of socially desirable responding across different survey modes. In reviewing the evidence across all types of questions, Schwarz and Strack (2003) propose that socially desirable responding is more likely to influence results in face-to-face interviews, then telephone interviews, and it is least likely to occur in confidential self-administered questionnaires. However, other authors have reported an *increase* in socially desirable responding to socially sensitive questions in telephone interviews, as compared with face-to-face methods (Holbrook, Green and Krosnick, 2003; Jäckle, Roberts and Lynn, 2006; Pudney, 2010). Some of the variability in findings may be due to within-mode variance, such as the various methods by which interviews can be conducted even if the overall modality is similar (e.g. with or without showcards; with or without some self-administered sections; computer-assisted versus pen-and-paper; randomised versus fixed presentation of the questions, etc.).

Evidence regarding mode-related social desirability and subjective well-being has to date been focused on evaluative measures and thus far contains quite mixed findings. Scherpenzeel and Eichenberger (2001) compared CATI and CAPI in a repeated-measures design (*N* = around 450) using questions drawn from the Swiss Household Panel Survey. No significant mode effects were found for life satisfaction – and the presence of the respondents' partner in one-third of the CAPI interviews also failed to influence responses. This finding is supported by Jäckle, Roberts and Lynn (2006), who conducted an experimental study in Hungary (*N* = 1 920) to examine the potential implications of shifting the European Social Survey (ESS) from the face-to-face procedure used currently to a telephone-based interview. Jäckle et al. also found no significant mode effects on mean scores of life satisfaction, even though some of the other socially sensitive questions tested did exhibit mode effects (for example, significantly higher household incomes were reported in the telephone condition).

In contrast with the above findings, Pudney (2010) reported that the survey mode (CAPI, CASI and CATI) had a significant influence on the distribution of responses across several domains of satisfaction in an experimental sub-panel of the UK Understanding Society survey (*N* = over 1 500). Among female respondents, there was a significant effect of mode on the distribution of scores on overall life satisfaction, overall job satisfaction and satisfaction with leisure time, and among male respondents a significant (CASI versus CAPI) difference in overall life satisfaction. Both CASI and CAPI tended to be associated with *lower* overall mean satisfaction levels in comparison to the telephone interview technique. Across all satisfaction domains other than income, CATI telephone interviewing also increased the likelihood that respondents would indicate that they were *completely* or *mostly* satisfied.

Pudney (2010) also found some evidence to suggest that the survey mode influenced the statistical relationships between satisfaction domains, individual characteristics and life circumstances. The results varied between different satisfaction domains, but there were some significant findings, the most notable being that the survey mode had a sizeable impact on the strength of the relationship between health satisfaction and two self-reported health predictors.[15] Although patchy, Pudney's results are important, because they imply that "the relationship between wellbeing and personal circumstances can be affected in important ways by apparently minor features of survey design" (p. 19).

Consistent with Pudney (2010), Conti and Pudney (2011) also found strong evidence of a mode effect on job satisfaction in a British Household Panel Survey data set. In this study, the same set of respondents completed both self-administered questionnaires and face-to-face interviews, administered consecutively in one visit to the respondent's home. Only 45% of respondents gave the same response in both the interview and the questionnaire, with a tendency for lower satisfaction reports in the questionnaire.

Although the influence of other survey context effects (such as adjacent questions, which differed between survey modes) cannot be ruled out, Conti and Pudney interpreted their results as being most consistent with self-presentation or social desirability effects influencing interview reporting. For example, the fact that having a partner present during the interview significantly depressed job satisfaction was regarded as being consistent with *strategic reporting behaviour, related to credibility and bargaining power within the family* – and specifically a "don't appear too satisfied in front of your partner" effect. The presence of children during the interview meanwhile made women more likely to report higher job satisfaction – a "not in front of the children" effect.

Conti and Pudney also found evidence of mode effects on the *determinants* of reported job satisfaction. One striking result is that, while in self-report questionnaire responses wages were an important determinant of job satisfaction for both men and women, the face-to-face interview data confirmed the typical finding that other non-wage job aspects were more important to women's job satisfaction. Women who worked longer hours were more likely to report lower job satisfaction in interview, but there was no significant association between hours worked and job satisfaction in the questionnaire report. The authors suggest that this implies female respondents were more likely to conform to social roles in the interview condition.

In a rare study looking at mode effects across a broader range of subjective well-being questions, the UK Office of National Statistics (2011b) recently tested the effect of survey mode on overall life satisfaction, a eudaimonia measure (*overall, to what extent do you feel the things you do in your life are worthwhile?*), happiness yesterday, and anxiety yesterday. In a national survey (N = 1 000), face-to-face interviews were contrasted with a laptop-based self-completion method. The only item that showed a statistically significant difference was anxiety yesterday (*overall, how anxious did you feel yesterday?*), where the mean average for the self-completion method was significantly higher than in the interviewer-led condition (mean = 3.7, compared to 3.2). Whilst it remains unclear what is driving this difference, social desirability effects are possible candidates. However, even in the self-completion condition, there was an interviewer present to administer the rest of the survey. It is thus possible that greater mode effects might be detected in the absence of an interviewer.

One other notable finding from the ONS work was that the self-completion condition had a much higher non-response rate than the face-to-face interviews (around 23%, compared to around 1.5%), and this was particularly marked among older participants. This implies that respondents might be *less* comfortable completing questions privately via a laptop. However, one difficulty in interpreting the finding is that *only* the subjective well-being questions were administered via a self-completion method: the remaining part of the longer interview was still conducted face-to-face. Thus, the subjective well-being questions were isolated as different from the others, which may have made respondents more nervous about completing them. This methodological feature also means that it is not possible to compare subjective well-being non-response rates with non-response rates for other items. The higher non-response rate for the laptop-administrated questions may reflect that the respondents (and especially the older respondents) did not wish to use the laptop in general, rather than that they have a particular aversion to completing subjective well-being questions via this method.

In summary, while several studies have suggested evidence of a significant social desirability mode effect on subjective well-being, others have failed to do so. Where effects do exist, the findings can be difficult to disentangle, and it is not always clear that the effects can really be attributed to socially desirable responding, rather than to other types of response biases.

Response biases and satisficing. There are a number of reasons to expect different survey modes to vary in their susceptibility to response biases and satisficing. The mode has implications for how respondents are contacted and motivated, and it also influences the level of burden associated with question and response formats. For example, the visual presentation of information in self-administered surveys (or interviews with showcards) can reduce the memory burden on respondents, which may in turn reduce satisficing. On

the other hand, visual presentation of text-based information places higher cognitive burdens on those with literacy problems, which is an important factor in obtaining representative samples where literacy rates are not universally high. For example, Jäckle et al. (2006) note that cross-cultural variations in literacy levels prohibit the sole use of self-administered questionnaires in the European Social Survey.

Some studies have suggested that telephone interviewing can lead to lower-quality data, relative to face-to-face interviews. For example, Jordan, Marcus and Reeder (1980) examined the impact of the survey mode on health attitudes among large US samples, and found that telephone interviewing induced greater response biases (acquiescence, evasiveness and extremeness) than face-to-face methods. Holbrook, Green and Krosnick (2003) meanwhile analysed satisficing, social desirability and respondent satisfaction in three carefully-selected large US data sets from 1976, 1982 and 2000. They replicated Jordan et al.'s findings of significantly greater response effects in telephone versus face-to-face interviews in lengthy surveys that examined issues such as political participation and attitudes.

In contrast, two more recent European studies failed to find evidence of greater satisficing among telephone-interview respondents when compared to face-to-face interviewees. Scherpenzeel and Eichenberger (2001) compared computer-assisted telephone and personal interview techniques (CATI and CAPI), using a selection of questions from the normal Swiss Household Panel Survey, on topics such as health, satisfaction, social networks, income, time budget and politics. They concluded that the "choice of CATI versus CAPI has no implications for the data quality, defined as validity and reliability" (p. 18). CATI was, however, cheaper to administer (SFR 47 per interview, contrasted with SFR 86 in CAPI) and enabled research to be completed more quickly.

The study by Jäckle, Roberts and Lynn (2006) described earlier also tested whether the use of showcards in face-to-face interviews affected data quality on socially sensitive items drawn from the European Social Survey. In general, they detected no differences in results obtained with and without showcards, implying that these questions had been successfully adapted for verbal-only presentation. Problems did arise, however, in adapting numerical questions about household income and hours watching television. The use of an open-ended format in the verbal channel but banded response categories in the visual channel resulted in large differences in means and response distributions, even though the topics addressed involved relatively more objective and behavioural measures.

Although self-administered pen-and-paper or web-based questionnaires may offer the greatest privacy for respondents (thus potentially reducing social desirability effects, Conti and Pudney, 2011), there is some evidence to suggest that they can lead to lower overall data quality, relative to interviewer-led methods. Kroh (2006) analysed evidence from the 2002 and 2003 waves of the German Socio-Economic Panel Study (N = 2 249) and found that the data quality for subjective well-being items presented in the auditory mode (CAPI and PAPI) was better overall than for pen-and-paper self-administered questionnaires. In a multi-trait, multi-method design, Kroh examined the amount of variance in three 11-point subjective well-being measures that could be attributed to method effects (i.e. measurement error) rather than the latent well-being factor. Across measures of life, health and income satisfaction, the method variance was consistently highest in the self-administered questionnaire mode. Reliability estimates for the health satisfaction measure were also significantly higher in CAPI, as compared to the self-administered questionnaire.[16]

Finally, there may be increased risk of day-of-week effects (see below) in self-administered pen-and-paper or web-based surveys if respondents choose particular times of the week to respond. For example, if respondents "save" this task for a weekend, that could have implications for affect measures in particular, which may be biased upwards due to an overall positive weekend effect on mood (Helliwell and Wang, 2011; Deaton 2011). This means that when using self-administered modes, it will be particularly important to record the exact date that the survey was completed to examine the risk of this effect in more detail.

Key messages on survey mode

From a data quality perspective, face-to-face interviewing appears to have many advantages. Interviewer-led techniques do, however, appear to be at slightly greater risk of prompting respondents to make more positive self-reports, relative to self-completion methods – and the finding that this tendency may be exacerbated in telephone surveys indicates that the potential privacy benefits of the telephone method could be outweighed by the additional rapport that interviews can establish in face-to-face conditions (Holbrook, Green and Krosnick, 2003). The presence of partners at the time of interview may also influence responses to some sensitive questions. Much of the evidence with regard to social desirability is ambiguous, however, and the significance of findings varies from study to study. There is relatively little evidence to suggest that subjective well-being measures are uniquely susceptible to mode effects, and in some cases, other social survey questions appear to be more sensitive. The evidence reviewed above suggests that in some cases the magnitude of mode effects on subjective well-being can be quite small, but where they do exist they may affect subsequent analyses of the data. The mixed findings in relation to the survey mode are likely to reflect both differences between studies in terms of the specific questions being asked, and also within-mode differences (e.g. with or without the use of show cards, with or without computer-assisted designs, etc.). It is also possible that the variation in results has arisen because some cultures may show stronger social desirability and audience effects than other cultures (see Section 5). As mixed-mode methods are increasingly being used for household surveys, our knowledge about the impact of the survey mode across a variety of cultures will continue to develop.

Given the impact that different question wording and response formats can have (discussed in Sections 1 and 2 of the current chapter), a critical issue for comparability where mixed-mode methods are used will be selecting questions that do not require extensive modification for presentation in different survey formats. Graphical scales (i.e. those that rely on visual presentation or a series of images, such as those with a series of faces from "smiley" to "sad"), very long question wording, and response formats that contain more than around five verbally-labelled response categories should in particular be avoided, because these will not translate well between different modes.

Where mixed-mode methods are employed, it will be essential to record details of the survey mode for each respondent and subsequently test for and report the presence or absence of survey mode effects. This will provide a larger pool of evidence that will enable more systematic examination of the role of the survey mode in subjective well-being data – including whether or not it may be possible to "correct" data for mode effects in the future.

The evidence described here focuses almost exclusively on evaluative subjective well-being. Much less is known about how mode effects could influence affective and eudaimonic measures. There is perhaps reason to expect that measures of recent affective experience would show a positive bias as a result of interviewer presence – it may be less easy to admit to a stranger (rather than record anonymously) that you've been feeling miserable lately. This is supported by the ONS (2011b) finding that self-completion respondents reported higher anxiety than those participating in face-to-face interviews. The pleasant mood induced by the positive social interaction of an interview experience may also influence retrospective recall for recent affect. These issues require empirical examination.

Wider survey context effects

The issue

Some concern has been raised that subjective well-being reports may be influenced by aspects of the wider survey context, such as the weather on the day of measurement, the day of the week that respondents were surveyed, and/or minor day-to-day events occurring immediately prior to the survey. While wider study context effects could be interpreted as a form of "satisficing" – where easily accessible information is used to help formulate responses, rather than necessarily the information that would lead to optimal answers – such effects could also indicate that respondents may simply find it difficult to distinguish between information sources when making subjective judgements or reporting subjective feelings.

Momentary mood is one particularly compelling information source that may interfere with longer-term evaluative judgements (Schwartz and Clore, 1983; Yardley and Rice, 1991) as well as retrospective recall processes (e.g. Bower, 1981; Clark and Teasdale, 1982; Herbert and Cohen, 1996). Retrospective judgements of emotional experience have also been found to display peak-end effects, whereby respondents give more weight to the most extreme experiences (the peaks) or to those experiences that have occurred most recently (the ends) (Redelmeier and Kahneman, 1996; Kahneman, 2003; Bolger, Davis and Rafaeli, 2003). This implies that, when making an assessment of subjective well-being, more recent experiences – such as those connected with very recent events – and more salient or extreme experiences – such as those that are subject to media focus – may disproportionately affect self-reports.

Two key measurement issues are at stake. The first is a threat to comparability: if different surveys are conducted in different contexts, does this limit the comparability of subjective well-being measures? And if so, what can be done to manage this? The second issue is the extent to which systematic context effects, particularly those capable of influencing a large proportion of respondents simultaneously, might drown out the impact of other significant and policy-relevant life circumstances that only affect a small number of respondents at any one time. This is perhaps more an issue of validity and data interpretation, rather than methodology, but there are nonetheless implications for the measurement approach adopted.

Day-to-day events. There is some evidence that short-term events can exert detectable effects on evaluations of subjective well-being. For example, Schwartz and Strack (2003) report some experimental manipulations in which finding a very small amount of money on a copy machine, spending time in a pleasant rather than unpleasant room, and watching a national football team win rather than lose a championship game served to increase

reported happiness and satisfaction with life as a whole. However, much of this evidence comes from small-scale studies with students, sometimes involving an impressive level of stage management.

A recent analysis has, however, highlighted the impact that a major news event and seasonal holidays can have on national-level reports of subjective well-being. Deaton (2011) examined a three-year time series of 1 000 daily subjective well-being reports in a national US representative sample, yielding close to 1 million respondents overall. Although the analysis focused on the effects of the financial crisis that began in summer 2008, some specific bumps in the time series data associated with short-term events are highlighted. For example, St. Valentine's Day produced a small one-day reduction in negative hedonic experiences (mean levels of *worry and stress, physical pain and anger*; and mean of *not happy, not enjoying, not smiling* and *sad*), and the Christmas holidays produced a much larger improvement in hedonic experience. There was also a sharp drop in overall life evaluations (measured using the Cantril Ladder) around the time of the collapse of Lehman Brothers in September 2008, which may have been due to respondents *anticipating* a potential change in their life circumstances as a result of this event.

Day of week. Day-of-week effects have been observed in large-scale national survey data. Taylor (2006) examined data from the 1992-2000 waves of the British Household Panel Survey (N = 8 301, contributing over 38 000 person-year observations) and found that both self-reported job satisfaction and subjective levels of mental distress systematically varied according to the day of the week when respondents were interviewed. In particular, controlling for a wide variety of other job-related, household and demographic determinants, men and women interviewed on Fridays and Saturdays reported significantly higher job satisfaction levels than those who were interviewed midweek, an effect that was particularly strong among full-time (as opposed to part-time) employees. In the case of mental distress, there were fewer significant effects, but employed women interviewed on Sundays reported significantly lower levels of mental well-being than those who were interviewed midweek (with an increase in stress levels of about 5% at the sample means).

Although the day of the week affected mean scores, Taylor found that the inclusion (or exclusion) of day-of-week controls did not alter observed relationships between job satisfaction or mental distress and job characteristics, education, demographics and employment status. However, the list of determinants investigated was not exhaustive, and there remains a risk that data collected on just one day of the week could systematically over- or under-estimate subjective well-being.

In another very large-scale data set, Helliwell and Wang (2011) found no day-of-week effects on life evaluations (Cantril Ladder), but a significant increase in *happiness, enjoyment* and *laughter*, and a significant decrease in *worry, sadness* and *anger* experienced on weekends and public holidays, relative to weekdays. This study examined data from 18 months of the Gallup Healthways Well-being daily telephone poll, in which 1 000 randomly-sampled adults in the United States are surveyed each day, yielding over half a million observations. Deaton (2011) reports the same pattern of results in a longer time series of the same data set, capturing nearly 1 million observations between January 2008 and December 2010.

Rather than calling into question the usefulness of subjective well-being data, the weekend effects observed by Helliwell and Wang are interpreted by the authors as evidence *in favour* of the validity of the measures used. In fact, one would *expect* that valid measures of momentary affect would vary according to the day of the week due to differences in

patterns of activity, whereas life evaluations should remain more stable over time. Consistent with this, the strength of the weekend effect observed by Helliwell and Wang varied in predictable ways according to job characteristics, such as being stronger for full-time employees relative to the rest of the population. A central variable in explaining the weekend effects was the amount of time spent with friends or family – which was on average 7.1 hours on a weekend, compared to 5.4 hours on week-days – and this increased social time at weekends "raises average happiness by about 2%".

Seasonal effects. Common folklore would certainly suggest a role for weather and climate in subjective well-being. This was demonstrated by Schkade and Kahneman (1998), who found that even though there were no significant differences between Californian and Midwestern US students in terms of their overall life satisfaction, respondents from *both* regions *expected* Californians to be more satisfied, and this expected difference was mediated by perceptions of climate-related aspects of life in the two regions.

The largest set of evidence about the role of seasonality in subjective well-being comes from clinically-depressed populations and those suffering from seasonal affective disorder (SAD); relatively less is known about the role of seasons on mood and positive mental states, such as life satisfaction, among normal population samples (Harmatz et al., 2000).

Harmatz et al. (2000) conducted a one-year longitudinal study in Massachusetts, US, taking repeated quarterly measures among a sample (N = 322) that excluded individuals positively assessed for SAD. Significant seasonal effects were present in the data, both on the Beck Depression Inventory, and on the emotion ratings for *anger, hostility, irritability* and *anxiety* scales, all of which followed the same seasonal pattern (highest in winter, lowest in summer, and spring and autumn somewhere in between these extremes). The general trend was evident in both male and female respondents, but was only significant for males in the cases of irritability and anxiety. Effect sizes were described by the authors as "relatively small" and "from a clinical perspective, these differences would be noted as mild mood fluctuations" (p. 349). However, mean score differences between summer and winter emotions for females were quite large, dropping at least one point on a 9-point scale between winter and summer.[17]

Seasonal patterns have also been detected among larger nationally-representative samples. Smith (1979) examined time trends in national subjective well-being data (N = 610 – 723 respondents per month) and found that reported happiness showed a seasonal pattern, with a 10 percentage-point range in the proportion of *very happy* respondents between the winter low and the spring high.[18] A positive affect measure also showed the same seasonal pattern (with a time series correlation with overall happiness of r = 0.83). On the other hand, a life satisfaction measure included in the study remained constant throughout the 12 months, and Bradburn's overall Affect Balance Scale also failed to show a seasonal effect. Curiously, negative affect actually dropped during the winter months, positively correlating (r = 0.66) with the happiness trend. This goes some way towards explaining the lack of a clear seasonal pattern in affect balance overall.

Smith (1979) also highlights the risk of drawing conclusions about seasonality based on only one year of data – pointing out that other context effects might be at play (e.g. the Arab oil embargo hit the US from October 1973 to March 1974, at the same time as the happiness study). Previous studies have also observed spring ups and winter downs

(Bradburn, 1969; Smith, 1979), but another national sample data set analysed by Smith, the US National Opinion Research Center's General Social Survey in 1972 (*N* = 1 599), failed to show a spring upswing.

The potential existence of seasonal effects raises the possibility that climate differences may account for some of the differences in subjective well-being observed between countries when objective life circumstances are controlled. This question was examined by Redhanz and Maddison (2005), using World Values Survey data on self-reported evaluations of happiness (on a 1-4 point scale), across a panel of 67 countries. Ten different climate indicators were constructed and tested in three different models, controlling for a range of other anticipated determinants of happiness (such as GDP per capita, unemployment, life expectancy and literacy). Three of the ten climate indicators had a significant effect: higher mean temperatures in the coldest month were found to increase happiness; higher mean temperatures in the hottest month were found to decrease happiness; and more months with very little precipitation were found to decrease happiness.

Seasonal effects may also be produced by seasonal trends in some of the key determinants of subjective well-being, such as unemployment status. Cycles in work and study situations may also be of relevance, particularly with regard to momentary mood. Whilst some of these effects will be substantive (i.e. they reflect how respondents actually feel, rather than contributing error to the data), they will nonetheless potentially have implications for how mean scores might vary over the course of a year, and therefore how data should be collected and described.

Weather. Evidence regarding the effects of weather on subjective well-being is mixed, and appears to be contingent on a number of factors. It has been suggested, for example, that although sunny days can produce higher estimates of both affect and life evaluations than rainy days, respondents might be able to exclude this information from their judgements if their attention is brought to it. Specifically, Schwarz and Clore (1983) found that among a very small sample of student telephone interviewees, respondents reported higher levels of current happiness, overall happiness with life, and satisfaction with life on sunny days as compared to rainy days. However, when the interviewer drew respondents' attention to the weather, there were no significant differences between evaluations on rainy versus sunny days.

The impact of cloud cover on life satisfaction has also been investigated. In two large-scale Canadian surveys (*N* = around 6 000 and 1 500) conducted over several months, Barrington-Leigh (2008) found that seven days of completely sunny weather more than doubled the chance of an individual reporting an extra point higher on a ten-point life satisfaction scale, as compared with a completely overcast week. This effect size was notable relative to other predictors examined.[19] However, including or excluding weather conditions in statistical estimates of life satisfaction from a range of other determinants (such as health, trust and income) produced indistinguishable coefficients.

In contrast to the work of both Schwarz and Clore and Barrington-Leigh, other evidence indicates no consistent effect of weather on life evaluations. Lawless and Lucas (2011) examined a large-scale survey data from over 1.5 million people over 5 years, and found no evidence of an effect of weather (rain, temperature, or combined weather conditions such as cold and rainy) on life satisfaction. The one exception to this was a significant interaction effect for *change* in weather conditions. Specifically, a temperature drop during warm months produced a very small increase in life satisfaction.

Although one might expect weather to have a significant impact on momentary mood, there is limited evidence of this, and the pattern of results is somewhat complex. Connolly Pray (2011) examined Princeton Affect and Time Survey data collected in the US *during summer months*, and found that among the women (but not the men) in their sample, low temperatures increased happiness and reduced tiredness and stress,[20] whereas higher temperatures reduced happiness. Despite observing a significant effect on life evaluations, Barrington-Leigh (2008) failed to find a significant relationship between cloud cover and a short-term affective measure of happiness. This makes Barrington-Leigh's results on life evaluations somewhat difficult to interpret, given that mood would be the primary mechanism through which one might expect weather to contaminate life evaluations.

In another large sample, Denissen et al. (2008) conducted an online repeated-measures diary study in Germany (N = 1 233) and found that six different weather parameters accounted for very little variance in day-to-day positive affect, negative affect or tiredness.[21] Importantly, however, there were individual differences in weather sensitivity, with the effects of hours of daylight varying the most between individuals – being, on average, 21 times greater than the average random effect of the other five weather variables. These individual differences in the relationship between weather and mood were not significantly associated with differences in age, gender or personality traits. It is possible that patterns of daily activity, such as the amount of time spent outside, could play a role (Keller et al., 2005).

Taken together, the results of Connolly Pray (2011), Lucas and Lawless (2011) and Barrington-Leigh (2008) tend to suggest that unusual weather events or shifts are most likely to have an impact on subjective well-being: living somewhere with year-round sunshine might mean a single day of sunshine has limited effects on activities, emotions or evaluations, whereas a rare day of sunshine during a long grey winter, or a rare day of cooler temperatures during a long hot summer, could have sufficient power to temporarily influence momentary affect and/or someone's outlook on life more generally. This suggests that the impact of short-term weather is most likely to be a problem in subjective well-being data when surveys are conducted on a single day, or on very small number of days, and within a limited geographic region. If a wider range of days and seasons are sampled, the effects of weather should be less problematic for mean levels of data aggregated across the study period.

Key messages on wider survey context effects

Time-, event- and weather-specific effects can be thought of as a source of *error* in life evaluation and eudaimonic measures, but they are primarily a source of *information* in the case of short-term affective measures. This further highlights the importance of the reference period associated with a given question, discussed in the *question construction* section above. Relatively ephemeral aspects of the environment can have important and valid impacts on momentary mood, but should not in theory show large impacts on long-term life evaluations. The evidence available bears this out to some extent, but there are exceptions – particularly in the case of major news stories or public holidays – and there is very little information available about the impact of wider survey context on eudaimonic measures. The fact that context can have valid impacts on momentary affect nonetheless has implications for its measurement – and these implications are in fact similar to those for managing the error that wider survey context might introduce in more evaluative subjective well-being measures.

Daily events that affect individuals more or less at random, or temporary weather conditions whose effects vary widely across samples for a variety of reasons, may contribute a small amount of noise to the data, but this should not substantially affect analysis of co-variates or the interpretation of group differences. There is very little that survey designers can do to control these events themselves, although where they may be of particular interest (for example, in time-use and other diary or experience sampling studies), useful information can be collected about their occurrence – such that their effects can be considered or controlled in the course of analyses.

On the other hand, major events, rare weather shifts or day or week effects that have the potential to influence a sizeable proportion of regional or national respondents can introduce more systematic impacts on subjective well-being. With some exceptions (such as those noted by Deaton, 2011, and Barrington-Leigh, 2008), effect sizes are not very large, but the comparability of data – between groups and over time – can be threatened. Comparable data will thus depend on consistency with regard to the proportion of weekday/weekend measurement, survey timing in relation to seasons, the inclusion/exclusion of holiday periods, and the absence of major news events or particularly good or bad runs of weather. Measurements restricted to a single day or single survey week may be particularly vulnerable to instability due to the risk of coinciding with significant events, holidays, religious festivals, etc. It is thus preferable to stage data collection over multiple days wherever possible – and ideally throughout the year. Deaton's (2011) work also highlights the value of exploring data in time series to check for sudden movements in the data that may need to be explored to determine whether they result from something of direct interest to survey data users.

There is also a concern that the hopes and fears that might arise as a result of watching stock market trends and media reporting, because of their cross-national effects on national "mood", may obscure the effects of policy-relevant life circumstances (such as unemployment), simply because a smaller number of people are affected in the latter condition. This has important implications for analysis and interpretation, but cannot be easily addressed through methodology alone. The one exception is perhaps the reference period emphasised to respondents. If no reference period is specified for life evaluations and eudaimonia, respondents may be more likely to draw on information from recent events, rather than actively searching their memories over a longer period. However, as noted in the earlier section on question construction, relatively little is known about how the reference period alters responses to these types of subjective well-being data. When it comes to recently-experienced affect, shorter reference periods are expected to be associated with more accurate recall – and there is therefore a trade-off to manage between response accuracy and the risk of the measurement day exerting undue influence. Where recent affect questions are to be included, it is therefore strongly advisable that these measurements are spread over a wide period of time, rather than focusing on a particular day or week of the year.

In analyses of panel data, the impact of seasonal trends as well as of relatively ephemeral events will need to be considered, in addition to changes in life circumstances that may drive changes in subjective well-being over time.

5. Response styles and the cultural context

Introduction

Response biases and heuristics have been cross-cutting themes throughout this chapter, with an emphasis on how survey methodology can affect their likelihood. As noted in the introduction, however, the risk of response biases, heuristics and error is essentially the product of a complex interaction between methodological factors (such as the cognitive demands made by certain questions), respondent factors (such as motivation, fatigue and memory) and the construct of interest itself (such as how interesting or relevant respondents find it).

Where present, response biases can affect the accuracy of self-report survey data. The precise nature of the effect depends, however, on the bias in question, what the bias has been caused by, and whether it is affecting all respondents and all items in a survey similarly and consistently over time. Some of these various possible biases, sources and impacts are summarised and illustrated in Table 2.3. Key risks highlighted include increased error in the data, biased mean scores (up or down), and risks to the overall accuracy of comparisons between groups, over time or between different surveys.

Table 2.3. **Illustrative examples of response biases and their possible effects**

Source of response bias or heuristic	Type of bias observed	Potential respondents affected	Potential impact on responses	Potential risks to further analyses	Discussed in the present chapter in…
Question wording or response format encourages acquiescence	Acquiescence.	Potentially all, but some may be more susceptible to acquiescence than others (e.g. those with lower motivation).	Responses biased in the direction of the positive response category.	• Risk of inflated associations with any other variables that use the same response scale. • Risk of less accurate comparisons between groups if some groups are *consistently* more susceptible than others.	Sections 1 and 2.
Prior survey questions prime respondents to think about certain information when responding	Priming (question order effects).	Potentially all, but some may be more susceptible to context effects than others (e.g. those with lower motivation).	Responses biased in the direction of the prime.	• Risk of less accurate comparisons between surveys with different question ordering. • Risk of inflated associations between the variable and the prime. • Risk of less accurate comparisons between groups if some groups are *consistently* more susceptible to context effects than others.	Section 3.
Sample includes some fatigued or unmotivated respondents	Satisficing (respondents more likely to exhibit response biases or use heuristics).	Only those experiencing fatigue/lower motivation.	Various, depending on the response bias or heuristic used by respondent.	• Random (largely unpredictable) error introduced. • Risk of less accurate comparisons between surveys (e.g. if respondents less fatigued by other survey methods). • Risk of less accurate comparisons over time if respondents less fatigued on subsequent occasion.	Sections 1, 2, 3 and 4.
Linguistic or cultural response "styles" (e.g. towards more moderate or more extreme responding)	Moderate responding/ extreme responding.	Different respondents affected in different ways (depending on language/culture).	Responses biased towards centre of response scale (for moderate responding) or towards extremes of response scale (for extreme responding).	• Risk of less accurate comparisons between groups, based on language or culture. • Minimal risk to comparisons over time and between surveys.	Section 5.

However, not all types of bias are a problem for all types of analyses, thus the management of response bias depends on both the nature of the bias and the nature of the analysis being performed.

Sections 1 and 2 of the current chapter explored the ways in which question wording and response formats can contribute to communication, motivation and memory failures – each of which are thought to present risks to the quality of subjective well-being data by making response biases and the reliance on response heuristics more likely. Section 3 meanwhile considered the extent to which priming and question context effects can influence patterns among subjective well-being data. Finally, in the course of examining mode effects, Section 4 discussed the extent to which subjective well-being measures may be affected by socially desirable responding.

In addition to the various methodological features that can influence response biases, it has been suggested that respondents themselves can also exhibit consistent *response styles* – i.e. a repeated tendency to rely on a particular response heuristic or show a relatively constant susceptibility to certain forms of response bias. This final section of the chapter explores the evidence in relation to response styles and subjective well-being in general, and the latter half of the section focuses in particular on the risk of *cultural* response styles that might affect the comparability of data between countries.

Response styles and shared method variance

The issue

If respondents exhibit habitual response styles when answering self-reported survey questions, this can present risks to the accuracy of the responses and any subsequent analyses that explore relationships between variables. One of the key risks associated with response styles is that, by introducing a relatively stable bias across several self-reported variables, they can artificially inflate correlations between those variables – a phenomenon often described in the literature as *shared* or *common method variance*. This is a particular problem for cross-sectional analyses of survey data. With longitudinal or panel data, it is possible to eliminate the impact of stable response styles (i.e. fixed effects) by focusing instead on which variables are able to predict *changes* in responses over time.

The section below briefly reviews the relatively rare studies that have attempted to actually measure and quantify specific response styles and considers some of the evidence regarding the *extent* to which response styles present a particular problem for subjective well-being data. Some illustrations of the problem of shared method variance are also provided. Implications for data analysis and interpretation are discussed at greater length in Chapter 4.

The evidence

In practice, it is almost impossible to say with certainty that a respondent's answers have been influenced by a response style – the presence of which is usually inferred from patterns observed across a range of survey items, rather than being externally verified against a common standard or actual behaviour. For example, one oft-used technique for measuring acquiescence in balanced multi-item scales (i.e. scales that have equal numbers of positively- and negatively-framed items[22]) involves simply adding up the scores on all items, without reverse scoring the negatively-framed items. The resulting figure could be regarded as a measure of the tendency to agree with the statements in the questions,

regardless of their meaning. However, this is a very indirect measure of acquiescence, and it is also inextricably bound to what could be genuine differences in scores on the variable of interest. Similarly, Marín, Gamba and Marín (1992) estimated acquiescence through counting the number of times a respondent agreed with a question and created an extreme responding indicator by counting the number of times a respondent chose either of the scale anchors (e.g. 1 or 5 on a 5-point Likert scale).

The available evidence on response styles and subjective well-being presents a mixed picture. In an interview-based study, Ross and Mirowsky (1984) examined the extent to which respondents generally exhibited acquiescent response tendencies across a range of survey items. They then examined the correlation between acquiescent tendencies and a number of other survey variables to see which ones appeared to be particularly affected by this response style. They failed to find a relationship between acquiescent tendencies and the reporting of symptoms of psychological distress, and controlling for acquiescence did not affect the relationship between socio-cultural variables and distress. In a pen-and-paper questionnaire-based study, however, Moum (1988) found a significant positive correlation between acquiescence and overall life satisfaction measured at three different points in time ($r = 0.14$, 0.14 and 0.18, $p < 0.01$, $N > 550$). Moum also found a significant relationship between the acquiescence measure and positive feelings in the last two weeks (*satisfied with self, life is worth living, in very good spirits*), but no significant relationship between acquiescence and negative feelings reported over the same time period (*lacked faith in self, life is meaningless, depressed*). Increased acquiescence was also found to be associated with increased age and lower education, although this did not influence the overall relationship between age, education and life satisfaction.

There has been a particular focus on *shared method variance* in the literature examining the relationship between positive and negative affect. As noted previously, some authors argue that these affective constructs are independent, i.e. minimally correlated, and others argue that they are polar opposites on a single dimension, which implies a strong negative correlation between them. If respondents have a tendency to adopt a particular pattern or response style, or fail to switch successfully between positively-framed items and negatively-framed items (as discussed in the section on *response formats*), this could reduce the strength of the negative correlation one might expect to see between positive and negative affect.

One set of studies, by Schimmack, Böckenholt and Reisenzein (2002), suggested that response styles have a negligible influence on affect measures. Using multi-trait, multi-method (MTMM) analyses, they examined the extent to which the correlation between positive affect (PA) and unpleasant affect (UA) could be explained by shared method variance (and hence, response styles or artefacts of the survey method). Re-analysing correlations from a series of different studies, they concluded that "none of the MTMM studies supported the hypothesis that response styles dramatically attenuate the correlation between PA and UA" (p. 468). In their own empirical work, they found no evidence of response styles producing positive correlations among subjective affect measures. However, in a judgement task involving more objective stimuli (six colours projected onto a screen, where participants had to rate the intensity of red, blue and yellow to various degrees), ratings of different colours were more closely related when the same response format was used. This implies that *other* types of questions may in fact be more susceptible to response styles than affect questions are.

In contrast, Watson and Clark (1997) reviewed two studies which found that random error, acquiescence and other systematic errors exert a "significant influence" on the relationship between positive and negative affect measures. Firstly, data from Tellegen et al. (1994, cited in Watson and Clark, 1997) showed raw correlations between positive affect and negative affect of r = -0.12 to r = -0.25; but controlling for random error increased this correlation to r = -0.28, and controlling for acquiescence raised the correlation to -0.43. Thus, scales that appeared largely independent on the basis of the raw data became moderately related after eliminating known sources of random and systematic error. Similarly, Diener, Smith and Fujita (1995, cited in Watson and Clark, 1997) found raw correlations between positive and negative mood ranging from r = 0.04 to r = -0.28, but after controlling for error through structural equation modelling, this rose to a latent correlation of r = -0.44. This goes some way towards supporting the concern of Green, Goldman and Salovey (1993) that true negative correlations between affect measures may be masked by response styles. However, Watson and Clark are keen to reiterate that "error is a universal problem in assessment, and that there is no evidence that self-rated affect is especially susceptible to either random or systematic error. Moreover, an accumulating body of data clearly supports the construct validity of self-rated affect" (p. 289).

According to Watson and Tellegen (2002), one particular type of affect measure is especially vulnerable to acquiescence bias – namely, repeated daily or within-day measures, aggregated over time. This aggregation technique is expected to reduce the impact of random measurement error, because random errors (uncorrelated across assessments) cancel each other out when observations are aggregated. However, Watson and Tellegen point out that if there are *systematic errors* in the data, the proportion of variance they explain in aggregated measures would increase, because systematic errors *are* correlated across assessments. For example, examining daily diary data (N = 392), Watson and Tellegen showed that their acquiescence measure shared only 4.8% of the variance with their measure of guilt from a single day, but 11% of the variance when guilt measures were aggregated over 10 days, and 16% of the variance when aggregated over 30 study days. The authors suggest that the impact of acquiescence increased over the number of aggregations in part because acquiescence was more stable than the emotion measure itself. Thus, the authors "strongly recommend that future researchers include acquiescence measures when repeatedly assessing mood" (p. 596).

The impact of response styles on eudaimonia is not clear, but some early work indicates potential cause for concern. In an interview-based study with a large US probability sample, Gove and Geerken (1977) found that nay-sayers reported significantly lower levels of generalised positive affect, and both nay-sayers and yea-sayers reported lower self-esteem than those who exhibited neither tendency. Their findings in relation to self-esteem may be relevant to the broader concept of eudaimonia.

As noted in Section 2, there may also be *a priori* grounds to expect a slightly heightened risk of acquiescence or socially desirable responding in eudaimonia measures due to the response formats frequently adopted on these measures. According to Krosnick (1999), the use of *agree/disagree, true/false* and, to a lesser extent, *yes/no* response formats are problematic because they are more susceptible to acquiescence bias. Two recently proposed measures of eudaimonia or psychological well-being[23] adopt a *strongly agree/strongly disagree* response format and an unbalanced set of scale items, with all, or all but one, items positively-keyed. There are thus perhaps *a priori* grounds to predict a heightened risk of acquiescence among these measures, and further research on this is warranted.

Another concern associated with response styles is that some groups of respondents may be more likely than others to exhibit them. Krosnick (1999) cites a number of studies indicating that response category order effects are stronger among respondents with lower cognitive skills. Gove and Geerken (1977) found that younger respondents were more likely to nay-say than older ones, and more highly educated respondents were also slightly more likely to nay-say. However, these differences did *not* appear to distort overall relationships between socio-demographic variables (including income, occupation, marital status, race, gender, age and education) and mental well-being. These findings suggest that where differences in subjective well-being are found between different age groups or between different educational groups, the possibility of response biases may be worth investigating alongside a range of other determinants.

Personality or temperament has also been linked to patterns of responses in relation to both subjective well-being and related constructs. One example is the role of "negative affectivity",[24] which has been associated with a more negative response style across a range of self-report questions, and which some authors (e.g. Burke, Brief and George, 1993; McCrae, 1990; Schaubroeck, Ganster and Fox, 1992; Spector, Zapf, Chen and Frese, 2000) therefore suggest should be controlled when performing cross-sectional analyses of self-report data.

Once again, however, finding a consistent pattern of more negative responding is not in itself proof of a negative "response style" that is adding *error* to the data, rather than proof of meaningful variation. The risks in controlling for personality or affectivity are twofold. First, controlling for personality could potentially swamp the effects of other important determinants and remove variance in subjective well-being data that is likely to be of policy interest. For example, if exposure to childhood poverty or long-term health problems influences responses to both personality and subjective well-being questions, controlling for personality in the analyses could mask the true impact of childhood poverty and/or long-term health problems on the outcomes of interest. Second, personality, and negative affectivity in particular, may also play a substantive role in the overall development of subjective well-being (e.g. Bolger and Zuckerman, 1995; Spector, Zapf, Chen and Frese, 2000). Thus, whilst it may be interesting to examine the role of personality and temperament in relation to subjective well-being where survey space enables both to be explored, it may not be advisable to "control for" personality in all analyses of subjective well-being co-variates.

Key messages on response styles

Response styles reflect "default" patterns of question-answering that respondents are particularly likely to rely on because they are unmotivated, fatigued, confused by the question, or unable to answer due to lack of knowledge or insight, or because of failures in memory. This has obvious implications for question construction, in that questions need to be as simple, easy to interpret and minimally burdensome as possible. It also reiterates that 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. This is true for all survey measures, and it is not clear from the present evidence that subjective well-being is at any greater risk of eliciting response styles than other self-reported survey items, especially socially sensitive ones.

If some groups of respondents are systematically more likely to exhibit response styles, this can threaten the accuracy of between-group comparisons in cross-sectional data. However, the literature in this area typically fails to clarify the extent to which

patterns observed in the data are simply due to response *style*, rather than a genuine difference in how life is experienced by those respondents – partly because the presence of response *styles* in subjective measures remains extremely difficult to detect with any certainty. These problems (and further discussion of potential solutions) are covered in more detail in the section on cultural differences in response styles, below.

Cultural differences in response styles and scale use

The issue

One particular concern raised in the literature is the extent to which individuals from different cultures or linguistic groups might exhibit systematically different response styles when answering subjective well-being questions. The presence of response styles or linguistic differences that systematically bias responses upwards, downwards or towards the most moderate response categories will distort scores and reduce the accuracy of comparisons between countries. As was the case with demographic and personality differences, it is, however, very difficult to separate differences in scale use and response styles from differences in the genuine subjective well-being of the different groups. This is a particular challenge for scales with subjective content, because unlike more "objective" reports (e.g. income), we lack the ability to cross-validate scores precisely against external references.

The evidence – wider literature

Some of the current evidence on cultural differences in response styles comes from beyond the subjective well-being literature. Marín, Gamba and Marín (1992) examined response styles among Hispanic and non-Hispanic White respondents in the USA, using a range of questions from four different surveys, each with an ordinal response scale, and concerning non-factual information (e.g. attitudes and beliefs rather than behavioural reporting). Their findings suggested that Hispanics preferred more extreme response categories and were more likely to agree with items (i.e. to acquiesce). However, the magnitude of difference varied substantially between studies: for example, in data set one, Hispanics reported extreme responses on 50% of items, whereas non-Hispanic Whites reported them on 47% of items (a very small difference); yet in data set three, Hispanics reported extreme responses on 72% of items, whereas non-Hispanic whites did so on only 58% of items (a much larger difference). Response patterns among more acculturated Hispanics were more similar to those of non-Hispanic whites.

Unfortunately, Marín, Gamba and Marín's study design (much like those adopted elsewhere in the literature) does not enable one to conclude that this pattern of responding represents greater *error*: Hispanics in this sample may have agreed more or reported more extreme responses because these best represent *how they actually feel*, not just how they report their feelings. Response styles are usually assumed to contribute error to the data – but where this is assumed, it is important to *demonstrate* this reduced accuracy or validity empirically. Almost no studies in the literature appear to take this extra step.

One exception can be found in the work of Van Herk, Poortinga and Verhallen (2004) who examined three sets of data from marketing studies in six EU countries (total $N > 6\,500$). They found systematic differences in acquiescence and extreme response styles, with both styles being more prevalent in data from Mediterranean countries (Greece, Italy and Spain) than from north-western Europe (Germany, France and the United Kingdom). Across 18 different sets of items, there were significant differences in the

extent of acquiescence across countries in 17 cases, and country differences had an average effect size of 0.074. Greek respondents were particularly high on extreme responding, and Spanish and Italian respondents also scored consistently higher than those from France, Germany and the United Kingdom. Country differences in extreme responding were significant in 12 out of 18 cases, with an effect size of 0.071, described as "almost of medium size". As the study also included measures of actual behaviour, the authors could be reasonably confident in attributing their results to response styles – in several tests, they failed to find a relationship between higher levels of scale endorsement and actual behaviour.

In contrast to the extreme response styles described above, it has been suggested that Asian Confucian cultures are more likely to show a preference for more moderate response categories (Cummins and Lau, 2010; Lau, Cummins and McPherson, 2005; Lee, Jones, Mineyama and Zhang, 2002) – although once again there is rarely empirical data demonstrating that this leads to less accurate or valid data. Lee et al. (2002), for example, reported that although culture (Japanese/Chinese/USA) did affect response patterns on a "sense of coherence" measure, this did not attenuate the observed relationship between "sense of coherence" and health – thus implying that scale validity was not adversely affected. What *did* seem to matter for scale validity in this study, however, was the number of Likert response categories used – and here, there was an interaction with culture, such that 7-point scales showed stronger relationships with health among Japanese respondents, whereas 4- and 5-point scales showed stronger relationships with health among Chinese and American respondents. As the authors themselves conclude, "this is rather disturbing and warrants further investigation" (p. 305).

Some researchers have also reported cultural differences in the extent to which socially desirable responding is likely. For example, Middleton and Jones (2000) found that small samples of undergraduate students from East Asian countries such as Hong Kong (China), Singapore, Thailand, Taiwan, Japan and China were more likely than North American students to report fewer socially undesirable traits and more socially desirable ones.

The evidence – subjective well-being

Acquiescence and extreme responding have also been investigated on constructs related to subjective well-being among a limited number of cultural and linguistic groups. Looking across measures that included psychological distress and locus of control (both loosely related to the construct of eudaimonia or psychological well-being), Ross and Mirowsky (1984) reported that Mexican respondents were more likely to exhibit acquiescent response tendencies, when compared with both Mexican-Americans and non-Hispanic Whites living in El Paso.

It has been hypothesised that certain cultures tend to use response scales differently, and that this could lie behind some of the subjective well-being differences observed between different countries with similar objective life circumstances. For example, Chen, Lee and Stevenson (1995, cited in Schimmack et al., 2002) suggested that low levels of life satisfaction in Japanese cultures may reflect a "modesty bias".

Tendencies either towards more extreme or more moderate responding could affect either the overall distribution of scores or the mean level of responses. Minkov (2009) explored differences among cultures in the extent to which strong and polarised (i.e. very good versus very bad) judgements tend to be reported across 17 different questions about

life quality judgements (e.g. satisfaction with income, family life and job) and social opinions (e.g. views on immigration, use of military force and protection of the environment).[25] Country scores for polarisation varied widely, with Middle Eastern Arab societies (such as Kuwait, Palestine territories, Egypt and Jordan) showing the greatest degree of polarisation in judgements, and East and South East Asian societies (such as Indonesia, Japan, China and Korea) the least polarisation.

Minkov's polarisation measure could reflect a variety of different effects, including differences among countries in the likelihood of using scale extremes, or the genuine diversity of social opinions within countries or greater differences in objective life circumstances within countries. Again, unfortunately, there is nothing in the data *per se* that enables one to separate these effects out – no attempt is made to demonstrate that the degree of polarisation is related to *error*. Minkov did, however, find evidence that polarisation was strongly correlated with other national-level dispositional and attitudinal differences – such as "dialecticism" (the tolerance for holding beliefs and feelings that may seem contradictory to a Western mind), the importance of demonstrating high levels of personal consistency, and the extent to which active-assertiveness or compliance are perceived as national traits.

Other studies have also shown how challenging it can be to separate out valid differences on a variable from the presence or absence of response styles. Of course, more moderate responding will produce more moderate mean scores, and more extreme responding will either influence the distribution of results (with a greater number of responses falling towards scale end-points) or potentially draw the mean value of a scale upward or downward where higher or lower scale values tend to be the dominant response within a group.[26] Illustrating this difficulty, Hamamura, Heine and Paulhus (2008) reported a clear tendency for North American students of European heritage to report higher self-esteem scores on the Rosenberg Self-Esteem Scale, as well as less "moderate responding" and less "ambivalent responding",[27] than either North Americans of East Asian origin or Japanese students (who had the lowest self-esteem and highest levels of response styles). However, across the whole sample (N = 4 835), there were strong and significant negative correlations between the self-esteem measure and moderate responding (r = -0.41); between the self-esteem measures and ambivalent responding (r = -0.65); and between self-esteem and East Asian ethnicity (r = -0.46).[28] The authors thus concluded that, "mean differences in self-esteem are inextricably confounded with these two response styles" (p. 940). This reflects how difficult it is to demonstrate that response styles add error to the data.

In a second study, Hamamura et al. (2008) attempted to isolate the effect of response style by focusing their attention on scale items where there were no overall mean score differences between cultures. This study involved a smaller sample (N = 185) of Canadian student volunteers of either European or East Asian heritage. Analyses were limited to 26 personality scale items, and once again these highlighted a relationship between East Asian ethnicity and heightened levels of moderate responding (r = 0.15, N = 185, $p < 0.05$) as well as more ambivalent responding (r = 0.18, N = 185, $p < 0.05$).

Further complicating the picture, Diener et al. (2000) have also found differences between countries in terms of their *ideal* level of satisfaction with life. They asked a sample of over 7 000 undergraduate students in 41 different societies to indicate the level of overall life satisfaction the "ideal person" would have. The variation was striking, ranging from a

mean average score of 19.8 (out of a possible 35) for China to 31.1 for Australia. Notably, those countries that typically score lower on evaluative measures reported relatively lower "ideal" scores (e.g. Japan, 25.8; Korea, 25.0; Hong Kong, China, 25.4), whereas some of those typically scoring higher on evaluative measures reported relatively higher "ideal" scores (e.g. Colombia, 31.0; Puerto Rico, 30.7). These differences between countries in the "ideal" level of life satisfaction could potentially influence the manner in which social desirability affects life evaluation data – with the most socially desirable response varying significantly between cultures.

Techniques based on item response theory have also been developed (e.g. Oishi, 2006; Vittersø, Biswas-Diener and Diener, 2005) to assist in the identification of differences in scale use (i.e. different patterns of number use). For example, Oishi (2006) found that Chinese and US respondents showed different patterns of item functioning on a set of life satisfaction questions, although this differential scale use only partly accounted for the mean differences between these respondent groups. Meanwhile, Vittersø, Biswas-Diener and Diener (2005) found that Greenlanders tended to use more extreme responding than Norwegians on the Satisfaction with Life Scale, but when this tendency was statistically controlled for Norwegians were found to be significantly more satisfied than Greenlanders.

Investigating national differences in patterns of responding can also help to identify subjective well-being questions that may not translate well between cultures, either linguistically or conceptually. For example, Vittersø et al. (2005) found that one item on the Satisfaction With Life Scale (*If I could live my life over, I would change almost nothing*) was particularly responsible for different response patterns between Norwegian and Greenlander respondents.

Chapter 4 discusses other techniques that have been used to explore the extent of cultural differences in scale use and the potential to adjust for their effects *post hoc*. This includes the use of counterfactuals (i.e. predicting levels of subjective well-being on the basis of observed determinants and comparing these estimates with the responses actually obtained); vignettes (short descriptions of hypothetical scenarios that respondents are asked to rate, and which may be used to identify differences in how respondents react to the same information); and migrant data (to test whether country-specific effects also apply to migrants in that country). These studies rarely focus on the identification of specific response styles *per se* (e.g. acquiescence or social desirability), but they seek to identify persistent upward or downward biases in how respondents from different countries self-report their own levels subjective well-being, which is very relevant to both moderate and extreme responding.

Key messages on cultural response styles and differences in scale use

Although there do appear to be some cultural or country differences in the patterns of responses observed across subjective well-being questions, very little is known about the extent to which this represents *error* in the data (rather than genuine differences in how people feel, or how they assess their lives). Perhaps the surest method for dealing with both individual and cultural variation in response styles is to adopt a repeated-measures panel survey design, which enables fixed effects to be controlled in subsequent analyses of the data. Analyses can then focus on examining the determinants of *change* in subjective well-being over time – which both side-steps the response style issue and offers the considerable advantage that causal relationships can begin to be explored.

Panel data do not, however, solve the problem of response styles potentially influencing the average levels of subjective well-being that might be reported by data providers such as national statistical agencies. Given the concerns around response styles and cultural bias, one option may be to focus international comparisons not on the *level* of responding, but (as in the analysis of panel data) on any *changes* in the pattern of responses over time (Cummins and Lau, 2010) – including on any differences in the *rate of change* between different population sub-groups over time. Internationally, then, the comparator of interest would be something like the percentage change in subjective well-being in different countries within a defined time period. There is already a precedent for this sort of approach in the reporting of GDP, where much of the headline reporting (such as the OECD's regular press releases) focuses on *GDP growth per quarter*, rather than on absolute levels of GDP between countries.

However, much as in the case of GDP, there will remain a strong desire to be able to compare average levels of subjective well-being between countries. Because of this, some authors have proposed methods to detect and adjust for national differences in scale use *post hoc*. These are discussed in more detail in Chapter 4.

Overall messages on response styles and scale use

The nature of subjective measures means that we can never really know whether one respondent's 8 out of 10 corresponds to the exact same mental state as another respondent's 8 out of 10. Individual differences in response styles and scale use may inevitably add some noise to self-report data – although the clear evidence for the validity of subjective well-being measures indicates that this noise is not a major threat to the usefulness of the data. The accuracy of comparisons between two groups of respondents may, however, be limited if it can be demonstrated that those two groups exhibit systematically different patterns of response styles or scale use.

Several empirical issues limit our ability to separate responses styles and differences in scale use from genuine real-score differences in the subject of interest. Applying *ex post* corrections, such as for the average number of agreements across the survey, prior to analysis might eliminate interesting sources of variation, introduce non-independence in the data and reduce interpretability. More sophisticated statistical techniques based on item response theory present a promising way forward in terms of identifying cultural differences in scale (i.e. number) use, but they do not eliminate the influence of several other types of response bias, such as acquiescence and social desirability (Oishi, 2006). We are also far from having reached firm conclusions in terms of what cultural differences in scale use mean for each country and how we should measure or correct for this in making international comparisons.

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 sound 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. Other sections of this chapter have discussed these issues in more detail. Of course, if the source of a response style is largely cultural, rather than related to the demands of the survey, it will be more difficult to address through question and survey design itself – and the methods for management include both collecting panel data and potentially applying *post hoc* adjustments (see Chapter 4).

Where a strong risk of fatigue-related response styles is anticipated, the length of the survey and the sequencing of questions should also be carefully considered – perhaps with more cognitively challenging questions timed to coincide with points in the survey where respondent motivation is likely to be highest. Of course, this needs to be balanced against recommendations from the previous section on question order (and in particular, the need to avoid contamination between subjective well-being and other socially sensitive survey items). As question order effects can sometimes be considerable, minimising these should be considered the first priority.

Although there is some evidence that response styles can influence responses to evaluative and affective questions, there is little reason to believe that subjective well-being measures are uniquely susceptible. In general, the effect of response styles on evaluative and affective responses appear to be small, and perhaps of most significance when examining cultural differences in reporting. Less is known about eudaimonic scales, however, and some scale design features may make them more vulnerable to response styles. In particular, there are *a priori* grounds to expect that questions with *agree-disagree* response formats might be more likely to elicit stronger acquiescence tendencies.

Overall conclusions and priorities for future work

This chapter has identified a wide range of methodological issues which have to be considered, and in some cases traded off, when attempting to measure subjective well-being through surveys in a way that produces high-quality and comparable data. Given the degree of sensitivity that subjective well-being measures show to varying survey conditions and to how questions are framed, guidelines for measurement perhaps need to be more precisely specified than is the case for some more "objective" indicators, such as oil production or life expectancy. Arguably, however, this sensitivity exists in many other self-reported measures, and thus guidelines should not need to be more rigorous than would be the case for several other social survey indicators – and particularly those requiring subjective judgments.

In terms of good practice for measuring subjective well-being, there are currently some known knowns, and these are summarised in the recommendations that follow. Several known unknowns remain, and these are the basis for the research priorities listed below. National statistical agencies are particularly well-placed to advance this research agenda, as they are in a unique position to undertake systematic methodological work with large and representative samples. Until more of this type of data has been collected, however, the best method for maximising data comparability, both within and between countries, will be to adopt a consistent approach across surveys. Recommendations on what this consistent approach should look like are captured in the draft survey modules provided in Chapter 3. Given the known sources of error in subjective well-being measures, further discussion of how to report, analyse and interpret this data is included in Chapter 4.

Question wording and response formats

Recommendations

- In terms of question design, wording obviously matters – and comparable measures require comparable wording. Effective translation procedures are therefore particularly important for international comparability.

- The length of the reference period is also critical for affect measures. From the perspective of obtaining accurate reports of affect actually *experienced*, reports over a period of around 24 hours or less are recommended. Evaluative and eudaimonic measures are intended to capture constructs spanning a much longer time period, but there is less evidence available regarding the ideal reference period to use.
- Variation in response formats can affect data quality and comparability – including between survey modes. In the case of evaluative measures, there is empirical support for the common practice of using 0-10 point numerical scales, anchored by verbal labels that represent conceptual absolutes (such as *completely satisfied/completely dissatisfied*). On balance, it seems preferable to label scale interval-points (between the anchors) with numerical, rather than verbal, labels, particularly for longer response scales.
- The order in which response categories are presented to respondents may be particularly important for telephone-based interviews and where each response category is given a verbal label. For numerical scales, this is likely to be less important, although consistent presentation of options from lowest (e.g. 0) to highest (e.g. 10) may be helpful in reducing respondent burden.
- In the case of affect measures, unipolar scales (i.e. those reflecting a continuous scale focused on only one dimension – such as those anchored from *never/not at all* through to *all the time/completely*) are desirable, as there are advantages to measuring positive and negative affect separately.
- For life evaluations and eudaimonia, there is less evidence on scale polarity. What information is available suggests that bipolar and unipolar measure produce very similar results for life evaluation measures, but bipolar scales may be confusing for respondents when evaluative questions are negatively-framed.

Priorities for future work

- One priority for future research is establishing whether valid and reliable short- or single-item measures can be developed for positive and negative affect and for eudaimonia measures.
- Establishing the optimal response format for affective measures – including the number of response options and whether frequency, intensity or binary scales should be preferred – is a key issue for future work. This should be linked to the quest to find the response format that can best convey scale unipolarity in a consistent manner to respondents. The most appropriate reference period for life satisfaction and eudaimonia measures also requires further work – and one key criterion in this case will be the policy relevance of the resulting measures.
- Further systematic investigation of the specific impact of response formats on response biases is also warranted, especially the relationship between *agree/disagree, true/false* and *yes/no* response formats and acquiescence.
- Additional research is also needed to better understand and prevent any problems respondents have in switching between positively- and negatively-framed questions. Across a module of several subjective well-being questions, it would also be helpful to know whether the benefits of keeping the response format the same (in terms of time and respondent effort) are greater than the benefits of changing response formats between questions (which could help to more clearly mark a difference between different measures, and could potentially also help respondents perform the mental switch required to answer different questions).

Question order and context effects

Recommendations

- Question order effects can be a significant problem, but one that can largely be managed when it is possible to ask subjective well-being questions before other sensitive survey items, allowing some distance between them. Where this is not possible, introductory text and other questions can also serve to buffer the impact of context.
- Order effects are also known to exist *within* sets of subjective well-being questions. Evidence suggests that question modules should include only one primary evaluative measure, flow from the general to the specific, and be consistent in the ordering of positive and negative affect measures (due to the risk that asking negative questions first may affect subsequent responses to positive questions, and vice versa).

Priorities for future work

- Further research should investigate the most effective introductory text and buffer items for minimising the impact of question order on subjective well-being responses.
- The trade-off between the advantages of randomising the presentation order of affect measures, and the potential error introduced by respondents switching between positive and negative items, needs to be further investigated.
- More work is also needed to examine the impact that subjective well-being questions can have on responses to *subsequent* self-reported questions (such as subjective health or poverty measures). Order effects seem to be reduced when the most general questions are asked first, and more specific questions second – and this would favour placing very general subjective well-being questions (e.g. life evaluations) ahead of other more specific self-report questions. Ideally, some distance between subjective well-being questions and other sensitive items would also be built into the survey design.

Survey mode and timing

Recommendations

- The use of different survey modes can produce differences in subjective well-being data – although the significance and magnitude of the differences varies considerably from study to study due to the large number of variables that can influence mode effects. Given the number of trade-offs to be considered when selecting between survey modes, there is no one clear "winner" – although from a data quality perspective, face-to-face interviewing appears to have a number of advantages.
- Where mixed-mode surveys are unavoidable, it will be important for data comparability to select questions and response formats that do not require extensive modifications for presentation in different modalities. Details of the survey mode should be recorded alongside responses, and mode effects across the data should be systematically tested and reported.
- Aspects of the wider survey context, such as the day of the week that the survey is conducted and day-to-day events occurring around the time of the survey, can influence short-term affective measures but this should not be regarded as error. However, there is also some evidence that rare and/or significant events can impact on life evaluations.

- In terms of methodological implications, the key concern is to ensure that a variety of days are sampled. Comparability of data can be supported through adoption of a consistent approach regarding the proportion of weekdays/weekends, holiday periods and seasons of the year sampled.

Priorities for future work

- As larger and higher-quality data sets become available on subjective well-being, the role of survey mode will become clearer – including any international differences in effects. It will be essential that mode effects across the data should be systematically tested and reported, enabling compilation of a more comprehensive inventory of questions known to be robust to mode effects.
- The effect of the wider survey context (day-to-day events, day of week, weather, climate, etc.) on eudaimonia remains largely unknown.

Response styles and international comparability

Recommendations

- Response styles present particular challenges for data interpretation when they vary systematically between countries or between population sub-groups within countries. This is relevant to all self-reported indicators, and there are not strong grounds for expecting subjective well-being measures to be uniquely affected.
- The best-known cure for response styles is currently prevention, through adopting sound survey design principles that minimise the risk that respondents will rely on characteristic response styles or heuristics to answer questions. This includes selecting questions that are easily translated and understood and minimally burdensome on memory, as well as structuring and introducing the survey in a way that promotes respondent motivation.

Priorities for future work

- Some of the factors influencing response biases and differences in scale use, such as respondent characteristics and the role of culture, cannot always be managed through good survey methodology alone. In particular, international and cultural differences in scale interpretation and use, which is linked to the broader conceptual translatability of subjective well-being content, may place limits on the comparability of data between countries.
- Systematic investigation of this issue presents a challenging research agenda. Until more is known about the international comparability of subjective well-being data, one option may be to focus reporting and analysis on *changes in subjective well-being over time*, including the use of panel data. This and other management strategies, including *post hoc* data adjustments, are discussed in Chapter 4.

Notes

1. For example, failures in memory can potentially be reduced by appropriate probing and time-marking (e.g. when responding to questions about affect experienced yesterday, respondents may benefit from a reminder such as "yesterday was tuesday"), whereas failures in communication may be reduced through appropriate translation procedures and pre-testing.

2. This may be particularly problematic in the context of official national household surveys, where respondents may find difficult to understand why government might want to collect information about just one day.

3. Correlations of $r = 0.62$ and 0.77 (both significant at the $p < 0.001$ level) were obtained between frequency judgements and experienced affect, whereas for intensity measures, correlations were $r = 0.54$ and $r = 0.59$ for positive affect intensity ($p < 0.001$); and just $r = 0.34$ and 0.13 for negative affect intensity ($p < 0.030$ and 0.23), respectively.

4. A large number of response options will only lead to greater scale sensitivity if respondents *actually use* all of the available options.

5. Respondents were asked to consider a shop or restaurant known to them and rate *overall quality* (extremely bad to extremely good) and a range of sub-categories, such as *competence of staff, promptness of service, range of choice*, etc. After completing the measures, respondents were then asked to rate the scales used in terms of: *ease of use, quick to use* and *allowed you to express your feelings adequately*.

6. Using a multi-trait, multi-method design, Kroh found validity estimates around 0.89 for the 11-point scale, 0.80 for the 7-point scale and 0.70 for the magnitude scale. In terms of reliability, the magnitude scale performed better, with 0.94 reliability across traits, compared to 0.83 for the 11-point scale and 0.79 for the 7-point measure. However, Kroh reports that the open-ended magnitude scale was generally more problematic, taking longer to complete and apparently reducing respondent motivation. Thus, on balance, and in particular due to the evidence on validity, Kroh recommends the use of 11-point scales.

7. Responses to the question: "In general, how happy are you with your life as a whole?".

8. In 1991, the scale anchors ranged from "not at all satisfied" to "completely satisified" – which could imply a unipolar scale. In 1992, this was changed to an unambiguously bipolar format: "completely dissatisfied" to "completely satisfied". This switch could also be partly responsible for the stronger skew in the 1992 and 1993 data: indicating complete dissatisfaction could be a much stronger statement than indicating the absence of satisfaction.

9. I.e. respondents who are less motivated, more fatigued or more cognitively burdened – and who may therefore be seeking the first satisfactory answer available, rather than giving detailed consideration to every response option.

10. *Cues* in this context refer to aspects of the survey that send signals to respondents about the information they may need to provide in their answers. For example, survey source can send a signal about the likely content of a survey, as well as the answers that might be expected.

11. The transition question used was: "Now thinking about your personal life, are you satisfied with your personal life today?" and the inclusion of this transition question reduced the impact of including political questions down from 0.6 of a rung to less than 0.1 of a rung.

12. There remains a (currently unexplored) risk that opening the survey with subjective well-being questions could then produce a context effect for other self-report and particularly subjective items, especially if the drive for consistency underpins some of these response patterns. Although there are *a priori* grounds to expect context to have less of an impact on more domain-specific judgements (Schwarz and Strack, 2003), it will still be important to investigate whether adding subjective well-being questions to the beginning of a survey has any detectable impact on other responses.

13. Schimmack and Oishi differentiate between *temporarily accessible* information, brought to mind as a result of study context (for example, item order), and *chronically accessible* information, which is available to individuals all the time and may be the default information used to generate overall life satisfaction judgments.

14. This implies assimilation when Negative Affect questions are asked first, but a contrast effect when Positive Affect questions are asked first. The ONS plan to run this test again to increase the sample size as well as to investigate order effects among single-item headline measures of evaluative, eudaimonic and affective subjective well-being. This will produce some helpful insights.

15. Contributing 21% to the variation in coefficient size when comparing CAPI and CASI survey methods.

16. Self-administered questionnaires consistently evidenced the highest amount of variance attributable to method effects (28%, 27% and 25% respectively), whereas the variance explained by method effects was much lower in the case of both PAPI (14%, 13%, 13%) and CAPI (11%, 9% and 10%). Reliability estimates for each of these satisfaction measures were broadly similar (e.g. for the life satisfaction scale, reliability was 0.79, 0.80 and 0.80 for SAQ, PAPI and CAPI respectively), with the exception of

health satisfaction (which varied from 0.79 in SAQ measures, to 0.83 in PAPI and 0.88 in CAPI). There was, however, an overall statistically significant difference in reliability between SAQ and CAPI, with CAPI producing higher estimates of reliability.

17. One challenge in interpreting this result, however, is that the study employed emotion rating scales that asked respondents to indicate their current seasonal level, *compared to how they generally feel*. This question phrasing may have encouraged respondents to reflect on seasonal and other contrasts.

18. Happiness was measured through a simple question, with only three response options: "Taking all things together, how would you say things are these days – would you say that you're very happy, pretty happy, or not too happy these days?"

19. For example, "The magnitude of the modelled effect of a change in weather circumstances from half-cloudy to completely sunny is comparable to that associated with more than a factor of ten increase in household income, more than a full-spectrum shift in perceived trust in neighbours, and nearly twice the entire benefit of being married as compared with being single" (p. 26).

20. These were short-term affect measures, in which respondents were asked to rate the intensity of feelings experienced the previous day on a 0-6 scale.

21. The authors found no effects of weather on positive affect, but small significant effects of temperature, wind and sunlight on negative affect – with warmer temperatures increasing, and both wind power and sunlight decreasing, negative affect. There was also a small but significant effect of sunlight on tiredness.

22. In this context, a "balanced" scale is a multiple-item scale that includes an equal number of positive- and negatively-framed questions – so, for example, an affect measure that contains equal numbers of positive and negative affect items. An "unbalanced" scale contains a disproportionate number of questions framed in a positive or negative way. So, for example, an unbalanced eudaimonia scale might have a greater number of positively-framed items (such as "Most days I get a sense of accomplishment from what I do"), relative to the number of negatively-framed items (such as "When things go wrong in my life it generally takes me a long time to get back to normal").

23. Diener and Biswas-Diener's (2009), *Psychological Well-Being Scale*; and Huppert and So's (2009), *Flourishing Index*.

24. I.e. a dispositional tendency towards experiencing negative affect.

25. Data were drawn from the Pew Global Attitudes Survey from 47 different nations across all continents (N = 45 239 interviews in 2007) and examined for the extent to which Likert scale extremes were used. Minkov's resulting "polarisation" measure was highest when 50% of respondents have chosen the positive extreme (e.g. very good), and 50% the negative extreme (e.g. very bad).

26. For example, a tendency for more extreme responding in a country where the majority of respondents score above the scale midpoint would result in a higher overall mean value because the positive extreme would be emphasised more often than the negative.

27. The ambivalence index constructed by these authors is described as capturing "the degree to which the respondent sees the true and false-key items as making opposite claims" (p. 935), and is a possible proxy for differences in dialectical thinking – i.e. the ability to tolerate holding what Western cultures might regard as contradictory beliefs.

28. All of these correlations were significant at the 0.01 level, two-tailed.

Bibliography

Alwin, D.F. (1997), "Feeling Thermometers versus 7-point Scales: Which are Better?", *Sociological Methods and Research*, Vol. 25, No. 3, pp. 318-340.

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.

Barrington-Leigh, C.P. (2008), "Weather as a Transient Influence on Survey-Reported Satisfaction with Life", *Munich Personal RePEc Archive (MPRA) Paper*, No. 25736, University of British Columbia, Vancouver, available online at: *http://mpra.ub.uni-muenchen.de/25736/*.

Benet-Martinez, V. and Z. Karakitapoglu-Aygün (2003), "The Interplay of Cultural Syndromes and Personality in Predicting Life Satisfaction: Comparing Asian Americans and European Americans", *Journal of Cross-Cultural Psychology*, Vol. 34, pp. 38-60.

Bishop, G. (1987), "Context Effects in Self Perceptions of Interests in Government and Public Affairs", in H.J. Hippler, N. Schwarz and S. Sudman (eds.), *Social Information Processing and Survey Methodology*, Springer Verlag, New York, pp. 179-199.

Bjørnskov, C. (2010), "How Comparable are the Gallup World Poll Life Satisfaction Data?", *Journal of Happiness Studies*, Vol. 11, pp. 41-60.

Blanton, H. and J. Jaccard (2006), "Arbitrary metrics in psychology", *American Psychologist*, Vol. 61, pp. 27-41.

Bolger, N., A. Davis and E. Rafaeli (2003), "Diary Methods: Capturing Life as it is Lived", *Annual, Review of Psychology*, Vol. 54, pp. 579-616.

Bolger, N. and A. Zuckerman (1995), "A framework for studying personality in the stress process", *Journal of personality and social psychology*, Vol. 69(5), p. 890.

Bower, G.H. (1981), "Mood and memory", *American Psychologist*, Vol. 36, No. 2, pp. 129-148.

Bradburn, N., 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.

Burke, M.J., A.P. Brief and J.M. George (1993), "The Role of Negative Affectivity in Understanding Relations between Self-Reports of Stressors and Strains: A Comment on the Applied Psychology Literature", *Journal of Applied Psychology*, Vol. 78, No. 3, pp. 402-412.

Campanelli, P. (2008), "Testing Survey Questions", in E.D. de Leeuw, J.J. Hox and D.A. Dillman (eds.), *International Handbook of Survey Methodology*, Larence Erlbaum, New York.

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.

Clark, A.E. and C. Senik (2011), "Is Happiness Different from Flourishing? Cross-Country Evidence from the ESS", *Revue d'Économie Politique*, Vol. 121, No. 1, pp. 17-34.

Clark, D.M. and J.D. Teasdale (1982), "Diurnal Variation in Clinical Depression and Accessibility of Memories of Positive and Negative Experiences", *Journal of Abnormal Psychology*, Vol. 91, No. 2, pp. 87-95.

Connolly Pray, M. (2011), "Some Like it Mild and Not Too Wet: The Influence of Weather on Subjective Well-Being", *Working Paper*, No. 11-16, Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi (CIRPÉE), Montreal.

Conti, G. and S. Pudney (2011), "Survey Design and the Analysis of Satisfaction", *The Review of Economics and Statistics*, Vol. 93, No. 3, pp. 1087-1093.

Costa, P.T. Jr. and R.R. McCrae (1987), "Neuroticism, Somatic Complaints and Disease: Is the Bark Worse than the Bite?", *Journal of Personality*, Vol. 55, No. 2, pp. 299-316.

Csikszentmihalyi, M. and R. Larson (1992), "Validity and Reliability of the Experience Sampling Method", in M.W. deVries (eds.), *The Experience of Psychopathology: Investigating Mental Disorders in their Natural Settings*, Cambridge University Press, New York.

Cummins, T. 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 National University, pp. 74-93.

Cummins, R.A. (2003), "Normative Life Satisfaction: Measurement Issues and a Homeostatic Model", *Social Indicators Research*, Vol. 64, pp. 225-256.

Cummins, R.A., R. Eckersley, J. Pallant, J. van Vugt and R. Misajon (2003), "Developing a National Index of Subjective Wellbeing: The Australian Unity Wellbeing Index", *Social Indicators Research*, Vol. 64, pp. 159-190.

Cummins, R.A. and A.L.D. Lau (2010), "Well-being across cultures: Issues of measurement and the interpretation of data", in K.D. Keith (ed.), *Cross-Cultural Psychology: A Contemporary Reader*, pp. 365-379, New York: Wiley/Blackwell.

Davern, M.T. and R.A. Cummins (2006), "Is life dissatisfaction the opposite of life satisfaction?", *Australian Journal of Psychology*, Vol. 58, No. 1, pp. 1-7.

Davern, M.T., R.A. Cummins and M.A. Stokes (2007), "Subjective Wellbeing as an Affective-Cognitive Construct", *Journal of Happiness Studies*, Vol. 8, pp. 429-449.

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, available online at: *www.nber.org/papers/w17128*.

DEFRA (2011), *Life Satisfaction and other Measures of Wellbeing In England, 2007-2011*, Department for the Environment, Food and Rural Affairs.

Denissen, J.J.A., L. Butalid, L. Penke and M.A.G. van Aken (2008), "The Effects of Weather on Daily Mood: A Multilevel Approach", *Emotion*, Vol. 8, No. 5, pp. 662-667.

Diener, E. and R. Biswas-Diener (2009a), "Scale of Positive and Negative Experience (SPANE)", in E. Diener (ed.), *Assessing Well-Being: The Collected Works of Ed Diener*, Springer, Dordrecht, pp. 262-263.

Diener, E. and R. Biswas-Diener (2009b), "Psychological Well-Being Scale (PWB)", in E. Diener (ed.), *Assessing Well-Being: The Collected Works of Ed Diener*, Springer, Dordrecht, p. 263

Diener, E., R.A. Emmons, R.J. Larsen and S. Griffin (1985), "The Satisfaction with Life Scale", *Journal of Personality Assessment*, Vol. 49, pp. 71-75.

Diener, E., R. Inglehart and L. Tay (2012), "Theory and Validity of Life Satisfaction Scales", *Social Indicators Research*, published in an online first edition, 13 May.

Diener, E., D. Kahneman, R. Arora, J. Harter and W. Tov (2009), "Income's Differential Influence on Judgements of Life versus Affective Well-Being", in E. Diener (ed.), *Assessing Well-Being: The Collected Works of Ed Diener*, Springer, Dordrecht.

Diener, E., C.K. Napa Scollon, S. Oishi, V. Dzokoto and E.M. Suh (2000), "Positivity and the Construction of Life Satisfaction Judgments: Global Happiness is Not the Sum of its Parts", *Journal of Happiness Studies*, Vol. 1, pp. 159-176.

Diener, E., S. Oishi and R.E. Lucas (2003), "Personality, Culture, and Subjective Well-Being: Emotional and Cognitive Evaluations of Life", *Annual Review of Psychology,* Vol. 54, pp. 403-425.

Diener, E., E. Sandvik, W. Pavot and D. Gallagher (1991), "Response Artifacts in the Measurement of Subjective Well-Being", *Social Indicators Research*, Vol. 24, pp. 35-56.

Diener, E., E.M. Suh, R.E. Lucas and H.L. Smith (1999), "Subjective Well-Being: Three Decades of Progress", *Psychological Bulletin*, Vol. 125, No. 2, pp. 276-302.

Diener, E., D. Wirtz, R. Biswas-Diener, W. Tov, C. Kim-Prieto, D.W. Choi and S. Oishi (2009), "New Measures of Well-Being", in E. Diener (ed.), *Assessing Well-Being: The Collected Works of Ed Diener*, Springer, Dordrecht.

Dolnicar, S. and B. Grün (2009), "Does one size fit all? The suitability of answer formats for different constructs measured", *Australasian Marketing Journal*, Vol. 17, pp. 58-64.

Eckenrode, J. and N. Bolger (1995), "Daily and Within-Day Event Measurement", in S. Cohen, R.C. Kessler and L.U. Gordon (eds.), *Measuring Stress: A Guide for Health and Social Scientists*, Oxford University Press, Oxford.

Eid, M. and M. Zickar (2007), "Detecting Response Styles and Faking in Personality and Organizational Assessments by Mixed Rasch Models", in M. von Davier and C.H. Carstensen (eds.), *Multivariate and Mixture Distribution Rasch Models*, Statistics for Social and Behavioral Sciences Part III, pp. 255-270, Springer Science and Business Media, New York.

Fowler, F.J. and C. Cosenza (2008), "Writing Effective Questions", in E.D. de Leeuw, J.J. Hox and D.A. Dillman (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum, New York.

Ganster, D.C. and J. Schaubroeck (1991), "Work stress and employee health", *Journal of Management*, Vol. 17, No. 2, pp. 235-271.

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.

Green, D.P., S.L. Goldman and P. Salovey (1993), "Measurement Error Masks Bipolarity in Affect Ratings", *Journal of Personality and Social Psychology*, Vol. 64, pp. 1029-1041.

Grice, H.P. (1975), "Logic and Conversation", in H. Geirsson and M. Losonsky (eds.), *Readings in Language and Mind* (1996), Wiley-Blackwell, Oxford.

Haberstroh, S., D. Oyserman, N. Schwartz, U. Kühnen and L.J. Ji (2002), "Is the Interdependent Self More *Sensitive* to Question Context than the Independent Self? Self-Construal and the Observation of Conversational Norms", *Journal of Experimental Social Psychology*, Vol. 38, pp. 323-329.

Hamamura, T., S.J. Heine and D.L. Paulhus (2008), "Cultural Differences in Response Styles: The Role of Dialectical Thinking", *Personality and Individual Differences*, Vol. 44, pp. 932-942.

Harmatz, M.G., A.D. Well, C.E. Overtree, K.Y. Kawamura, M. Rosal and I.S. Ockene (2000), "Seasonal Variation of Depression and Other Mood: A Longitudinal Approach", *Journal of Biological Rhythms*, Vol. 15, No. 4, pp. 344-350.

Hedges, S.M., L. Jandorf and A.A. Stone (1985), "Meaning of Daily Mood Assessments", *Journal of Personality and Social Psychology*, Vol. 48, No. 2, pp. 428-434.

Helliwell, J.F., R. Layard and J. Sachs (2012), *World Happiness Report*, Earth Institute, Columbia University.

Helliwell, J.F. and R.D. Putnam (2004), "The Social Context of Well-Being", *Philosophical Transactions of the Royal Society*, London B, Vol. 359, pp. 1435-1446.

Helliwell, J.F. and S. Wang (2011) "Weekends and Subjective Well-Being", *Working Paper*, No. 17180, National Bureau of Economic Research, Cambridge MA.

Herbert, T.B. and S. Cohen (1996), "Measurement Issues in Research on Psychosocial Stress", in H.B. Kaplan (ed.), *Psychosocial Stress: Perspectives on Structure, Theory, Life-Course, and Methods*, Academic Press, Inc., San Diego, CA.

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.

Hormuth, S.E. (1986), "The Sampling of Experiences in situ", *Journal of Personality*, Vol. 54, pp. 262-293.

Huppert, F.A., N. Marks, A. Clark, J. Siegrist, A. Stutzer, J. Vittersø and M. Wahrendorf (2009), "Measuring Well-Being across Europe: Description of the ESS Well-Being Module and Preliminary Findings", *Social Indicators Research*, Vol. 91, pp. 301-315.

Huppert, F.A. and T.T.C. So (2011), "Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being", *Social Indicators Research*, published online first, 15 December, DOI: *http://dx.doi.org/10.1007/s11205-011-9966-7*.

Huppert, F.A. and T.T.C. So (2009), "What Percentage of People in Europe are Flourishing and What Characterises Them?", Well-Being Institute, University of Cambridge, mimeo prepared for the OECD/ISQOLS meeting on *Measuring subjective well-being: an opportunity for NGOs?*, Florence, 23/24 July, available online at: *www.isqols2009.istitutodeglinnocenti.it/Content_en/Huppert.pdf*.

Jäckle, A., C. Roberts and P. Lynn (2006), "Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project", *ISER Working Paper*, No. 2006-41, August, University of Essex, Colchester.

Jordan, L.A., A.C. Marcus and L.G. Reeder (1980), "Response Styles in Telephone and Household Interviewing: A Field Experiment", *Public Opinion Quarterly*, Vol. 44, No. 2, pp. 210-222.

Kahneman, D. (2003), "Objective Happiness", in D. Kahneman, E. Diener and N. Schwarz (eds.), *Well-being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, New York.

Kahneman, D., A.B. Krueger, D.A. Schkade, N. Schwarz and A.A. Stone (2004), "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method", *Science*, Vol. 306, No. 5702, pp. 1776-1780.

Keller, M.C., B.L. Fredrickson, O. Ybarra, S. Côte, K. Johnson, J. Mikels, A. Conway and T. Wager (2005), "A Warm Heart and a Clear Head: The Contingent Effects of Weather on Mood and Cognition", *Psychological Science*, Vol. 16, No. 9, pp. 724-731.

Kroh, M. (2006), "An Experimental Evaluation of Popular Well-Being Measures", *German Institute for Economic Research Discussion Paper*, No. 546, Berlin, January.

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.

Krosnick, J.A. and M.K. Berent (1993), "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Questions Format", *American Journal of Political Science*, Vol. 37, No. 3, pp. 941-964.

Krueger, A.B. and D.A. Schkade (2008), "The Reliability of Subjective Well-Being Measures", *Journal of Public Economics*, Vol. 92, No. 8-9, pp. 1833-1845.

Larson, R. and P.A.E.G. Delespaul (1992), "Analyzing Experience Sampling Data: A Guidebook for the Perplexed", in M. deVries (ed.), *The Experience of Psychopathology*, Cambridge University Press, New York.

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.

Lawless, N.M. and R.E. Lucas (2011), "Predictors of Regional Well-Being: A County-Level Analysis", *Social Indicators Research*, Vol. 101, pp. 341-357.

Lee, J.W., P.S. Jones, Y. Mineyama and X. E. Zhang (2002), "Cultural differences in responses to a Likert scale", *Research in Nursing and Health*, Vol. 25, pp. 295-306.

Lim, H.E. (2008), "The Use of Different Happiness Rating Scales: Bias and Comparison Problem?", *Social Indicators Research*, Vol. 87, pp. 259-267.

Linley, P.A., J. Maltby, A.M. Wood, G. Osborne and R. Hurling (2009), "Measuring Happiness: The Higher Order Factor Structure of Subjective and Psychological Well-Being Measures", *Personality and Individual Differences*, Vol. 47, pp. 878-884.

Lucas, R.E. and M.B. Donnellan (2011), "Estimating the Reliability of Single-Item Life Satisfaction Measures: Results from Four National Panel Studies", *Social Indicators Research*, in press, available online first, 13 January, DOI: *http://dx.doi.org10.1007/s11205-011-9783-z*.

Lucas, R.E. and N.M. Lawless (2011), "Weather Conditions are Unrelated to Life Satisfaction Judgments: Evidence from a Large Representative Sample in the US", paper submitted for publication, Michigan State University.

Lynn, R. and T. Martin (1997), "Gender Differences in Extraversion, Neuroticism, and Psychoticism in 37 Nations", *The Journal of Social Psychology*, Vol. 137, No. 3, pp. 369-373.

Maggino, F. (2009), "Methodological aspects and technical approaches in measuring subjective well-being", Università degli Studi di Firenze, *Working Paper*, annexed to F. Maggino, "The state of the art in indicators construction", also a Università degli Studi di Firenze *Working Paper*, available online at: *http://eprints.unifi.it/archive/00001984/*, last accessed 11 July 2012.

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.

McCrae, R.R. (1990), "Controlling Neuroticism in the Measurement of Stress", *Stress Medicine*, Vol. 6, pp. 237-241.

McCrae, R.R., A. Terracciano, A. Realo and J. Allik (2007), "Climatic Warmth and National Wealth: Some Culture-Level Determinants of National Character Stereotypes", *European Journal of Personality*, Vol. 21, pp. 953-976.

Michalos, A.C. and P.M. Kahlke (2010), "Stability and Sensitivity in Perceived Quality of Life Measures: Some Panel Results", *Social Indicators Research*, Vol. 98, pp. 403-434.

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.

Minkov, M. (2009), "Nations with More Dialectical Selves Exhibit Lower Polarization in Life Quality Judgments and Social Opinions", *Cross-Cultural Research*, Vol. 43, pp. 230-250.

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.

Norenzayan, A. and N. Schwarz (1999), "Telling What They Want to Know: Participants Tailor Causal Attributions to Researchers' Interests", *European Journal of Social Psychology*, Vol. 29, pp. 1011-1020, available online at: *http://hdl.handle.net/2027.42/34565*.

Office for National Statistics, UK (2011a), Response times for subjective well-being experimental question trials, included in the *Integrated Household Survey*, early Summer 2011, Personal communication.

Office for National Statistics, UK (2011b), "Initial Investigation into Subjective Well-Being from the Opinions Survey", *Working Paper*, released 1 December, ONS, Newport, available online at: *www.ons.gov.uk/ons/rel/wellbeing/measuring-subjective-wellbeing-in-the-uk/investigation-of-subjective-well-being-data-from-the-ons-opinions-survey/initial-investigation-into-subjective-well-being-from-the-opinions-survey.html*.

Office for National Statistics, UK (2011c), "Subjective well-being: A qualitative investigation of subjective well-being questions", results of cognitive testing carried out during development of the ONS's experimental subjective well-being questions, December, unpublished report, shared with OECD through personal communication.

Oishi, S. (2006), "The Concept of Life Satisfaction Across Cultures: An IRT Analysis", *Journal of Research in Personality*, Vol. 40, pp. 411-423.

Oishi, S., Schimmack, U. and S.J. Colcombe (2003), "The Contextual and Systematic Nature of Life Satisfaction Judgments", *Journal of Experimental Social Psychology*, Vol. 39, pp. 232-247.

Parkinson, B., R.B. Briner, S. Reynolds and P. Totterdell (1995), "Time Frames for Mood: Relations Between Momentary and Generalized Ratings of Affect", *Personality and Social Psychology Bulletin*, Vol. 21, No. 4, pp. 331-339.

Pavot, W. and E. Diener (1993a), "The Affective and Cognitive Context of Self-Reported Measures of Subjective Well-Being", *Social Indicators Research*, Vol. 28, pp. 1-20.

Pavot, W. and E. Diener (1993b), "Review of the Satisfaction With Life Scale", *Psychological Assessment*, Vol. 5, No. 2, pp. 164-172.

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.

Podsakoff, P.M., S.B. MacKenzie, J.Y. Lee and N.P. Podsakoff (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.

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 Paper*, No. 2010-20, University of Essex.

Rässler, S. and R.T. Riphahn (2006), "Survey item nonresponse and its treatment", *Allgemeines Statistisches Archiv*, Vol. 90, pp. 217-232.

Redelmeier, D.A. and D. Kahneman, (1996), "Patients' Memories of Painful Medical Treatments: Real-Time and Retrospective Evaluations of Two Minimally Invasive Procedures", *Pain*, Vol. 66, No. 1, pp. 3-8.

Redhanz, K. and D. Maddison (2005), "Climate and Happiness", *Ecological Economics*, Vol. 52, pp. 111-125.

Robins, R.W., H.M. Hendin and K.J. Trzesniewski (2001), "Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale", *Personality and Social Psychology Bulletin*, Vol. 27, pp. 151-161.

Robinson, M.D., E.C. Solberg, P.T. Vargas and M. Tamir (2003), "Trait as Default: Extraversion, Subjective Well-Being, and the Distinction Between Neutral and Positive Events", *Journal of Personality and Social Psychology*, Vol. 85, No. 3, pp. 517-527.

Ross, C.E. and J. Mirowsky (1984), "Socially-Desirable Response and Acquiescence in a Cross-Cultural Survey of Mental Health", *Journal of Health and Social Behaviour*, Vol. 25, pp. 189-197.

Russell, J.A. (1980) "A Circumplex Model of Affect", *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161-1178.

Russell, J.A. and J.M. Carroll (1999), "On the Bipolarity of Positive and Negative Affect", *Psychological Bulletin*, Vol. 125, No. 1, pp. 3-30.

Russell, J.A., M. Lewicka and T. Niit (1989), "A Cross-Cultural Study of a Circumplex Model of Affect", *Journal of Personality and Social Psychology*, Vol. 57, No. 5, pp. 848-856.

Russell, J.A., A. Weiss and G.A. Mendelsohn (1989), "Affect Grid: A Single-Item Measure of Pleasure and Arousal", *Journal of Personality and Social Psychology*, Vol. 57, No. 3, pp. 493-502.

Ryff, C.D. and C.L.M. Keyes (1995), "The Structure of Psychological Well-Being Revisited", *Journal of Personality and Social Psychology*, Vol. 69, No. 4, pp. 719-727.

Saris, W.E., T. Van Wijk and A. Scherpenzeel (1998), "Validity and Reliability of Subjective Social Indicators: The Effect of Different Measures of Association", *Social Indicators Research*, Vol. 45, pp. 173-199.

Schaeffer, N.C. (1991) "Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers", *Public Opinion Quarterly*, Vol. 55, pp. 395-423.

Schaubroeck, J., D.C. Ganster and M.L. Fox (1992). "Dispositional affect and work-related stress", *Journal of Applied Psychology*, Vol. 77(3), p. 322.

Scherpenzeel, A. (1999), "Why Use 11-point Scales?", *Swiss Household Panel Working Paper*, University of Lausanne.

Scherpenzeel, A. and P. Eichenberger (2001), "Mode Effects in Panel Surveys: A Comparison of CAPI and CATI", *Swiss Federal Office of Statistics Working Paper*, order No. 448-0100, October.

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.

Schimmack, U. and S. Oishi (2005), "The Influence of Chronically and Temporarily Accessible Information on Life Satisfaction Judgments", *Journal of Personality and Social Psychology*, Vol. 89, No. 3, pp. 395-406.

Schimmack, U., S. Oishi and E. Diener (2002), "Cultural Influences on the Relation Between Pleasant Emotions and Unpleasant Emotions: Asian Dialectic Philosophies or Individualism-Collectivism?", *Cognition and Emotion*, Vol. 16, No. 6, pp. 705-719.

Schimmack, U., P. Radhakrishnan, S. Oishi, V. Dzokoto and S. Ahadi (2002), "Culture, Personality, and Subjective Well-Being: Integrating Process Models of Life Satisfaction", *Journal of Personality and Social Psychology*, Vol. 82, No. 4, pp. 582-593.

Schimmack, U., J. Schupp and G.G. Wagner (2008), "The Influence of Environment and Personality on the Affective and Cognitive Component of Subjective Well-Being", *Social Indicators Research*, Vol. 89, pp. 41-60.

Schkade, D.A. and D. Kahneman (1998), "Does Living in California Make People Happy? A Focusing Illusion on Judgements of Life Satisfaction", *Psychological Science*, Vol. 9, No. 5, pp. 340-346.

Schober, M.F. and F.G. Conrad (1997), "Does Conversational Interviewing Reduce Survey Measurement Error?", *Public Opinion Quarterly*, Vol. 61, pp. 576-602.

Schuman, H. and S. Presser (1981), *Questions and answers in attitude surveys: Experiments in question form, wording and context*, Academic Press, New York.

Schwartz, J.E. and A.A. Stone (1998), "Strategies for Analyzing Ecological Momentary Assessment Data", *Health Psychology*, Vol. 17, No. 1, pp. 6-16.

Schwarz, N. (1999), "Self Reports: How Questions Shape the Answers", *American Psychology*, Vol. 54, No. 2, pp. 93-105.

Schwarz, N. and G.L. Clore (1983), "Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States", *Journal of Personality and Social Psychology*, Vol. 45, No. 3, pp. 513-523.

Schwarz, N., H.J. Hippler, B. Deutsch and F. Strack (1985), "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments", *Public Opinion Quarterly*, Vol. 49, pp. 388-395.

Schwarz, N., B. Knäuper, D. Oyserman and C. Stich (2008), "The Psychology of Asking Questions", in E.D. de Leeuw, J.J. Hox and D.A. Dillman (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, New York.

Schwarz, N., B. Knäuper, H.J. Hippler, E. Noelle-Neumann and L. Clark (1991), "Rating Scales' Numeric Values may Change the Meaning of Scale Labels", *Public Opinion Quarterly*, Vol. 55, No. 4, pp. 570-582.

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.

Schwartz, N. and F. Strack (2003), "Reports of Subjective Well-Being: Judgemental Processes and their Methodological Implications", in D. Kahneman, E. Diener and N. Schwartz (eds.), *Well-being: The foundations of hedonic psychology*, Russell Sage Foundation, New York.

Schwartz, N. and F. Strack (1991), "Evaluating One's Life: A Judgement Model of Subjective Well-Being", in F. Strack, M. Argyle and N. Schwartz (eds.), *Subjective Well-Being: An Interdisciplinary Perspective*, Pergamon Press, Oxford.

Schwarz, N., F. Strack and H. Mai (1991), "Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis", *Public Opinion Quarterly*, Vol. 55, pp. 3-23.

Scollon, C.N., C. Kim-Prieto and E. Diener (2003), "Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses", *Journal of Happiness Studies*, Vol. 4, pp. 5-34.

Segura, S.L. and V. González-Romá (2003), "How do respondents construe ambiguous response formats of affect items?", *Journal of Personality and Social Psychology*, Vol. 85, No. 5, pp. 956-968.

Sgroi, D., E. Proto, A.J. Oswald and A. Dobson (2010), "Priming and the Reliability of Subjective Well-Being Measures", *Warwick Economic Research Paper*, No. 935, University of Warwick, Department of Economics, available online at: *www2.warwick.ac.uk/fac/soc/economics/research/workingpapers/2010/twerp_935.pdf*.

Simonsohn, U. (2007), "Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making", *Journal of Behavioural Decision Making*, Vol. 20, No. 2, pp. 143-152.

Smith, C. (2013), "Making Happiness Count: Four Myths about Subjective Measures of Well-Being", *OECD Paper*, prepared for the ISI 2011: Special Topic Session 26.

Smith, P.B. (2004), "Acquiescent Response Bias as an Aspect of Cultural Communication Style", *Journal of Cross-Cultural Psychology*, Vol. 35, No. 1, pp. 50-61.

Smith, T. (1982), "Conditional Order Effects", *General Social Survey Technical Report*, No. 33, NORC, Chicago.

Smith, T.W. (1979), "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences and Other Mysteries", *Social Psychology Quarterly*, Vol. 42, pp. 18-30.

Smith, D.M., N. Schwartz, T.A. Roberts and P.A. Ubel (2006), "Why are You Calling Me? How Study Introductions Change Response Patterns", *Quality of Life Research*, Vol. 15, pp. 621-630.

Smyth, J.M. and A.A. Stone (2003), "Ecological Momentary Assessment Research in Behavioral Medicine", *Journal of Happiness Studies*, Vol. 4, pp. 35-52.

Spector, P.E., D. Zapf, P.Y. Chen and M. Frese (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(1), pp. 79-95.

Stone, A.A. (1995), "Measurement of Affective Response", in S. Cohen, R.C. Kessler and L.U. Gordon (eds.), *Measuring Stress: A Guide for Health and Social Scientists*, Oxford University Press, Oxford.

Stone, A.A., J.E. Schwartz, J.M. Neale, S. Shiffman, C.A. Marco, M. Hickcox, J. Paty, L.S. Porter and L.J. Cruise (1998), "A Comparison of Coping Assessed by Ecological Momentary Assessment and Retrospective Recall", *Journal of Personality and Social Psychology*, Vol. 74, No. 6, pp. 1670-1680.

Stone, A.A., S.S. Shiffman, J.E. Schwartz, J.E. Broderick and M.R. Hufford (2002), "Patient Non-Compliance with Paper Diaries", *British Medical Journal*, Vol. 324, pp. 1193-1194.

Strack, F., L. Martin and N. Schwarz (1988), "Priming and Communication: The Social Determinants of Information Use in Judgments of Life Satisfaction", *European Journal of Social Psychology*, Vol. 18, pp. 429-42.

Strack, F., N. Schwarz and E. Gschneidinger (1985), "Happiness and Reminiscing: The Role of Time Perspective, Affect, and Mode of Thinking", *Journal of Personality and Social Psychology*, Vol. 49, No. 6, pp. 1460-1469.

Strack, F., N. Schwartz and M. Wänke (1991), "Semantic and Pragmatic Aspects of Context Effects in Social and Psychological Research", *Social Cognition*, Vol. 1, pp. 111-125.

Sudman, S., N.M. Bradburn and N. Schwarz (1996), *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*, Jossey-Bass, San Francisco.

Suh, E.M., E. Diener and J.A. Updegraff (2008), "From Culture to Priming Conditions: Self-Construal Influences on Life Satisfaction Judgments", *Journal of Cross-Cultural Psychology*, Vol. 39, No. 1, pp. 3-15.

Suls, J. and R. Martin (2005), "The Daily Life of the Garden-Variety Neurotic: Reactivity, Stressor Exposure, Mood Spillover, and Maladaptive Coping", *Journal of Personality*, Vol. 73, No. 6, pp. 1485-1510.

Taylor, M.P. (2006), "Tell Me Why I Don't Like Mondays: Investigating Day of the Week Effects on Job Satisfaction and Psychological Well-Being", *Journal of the Royal Statistical Society Series* A, Vol. 169, No. 1, pp. 127-142.

Tellegen, A. (1985), "Structures of Mood and Personality and their Relevance to Assessing Anxiety, with an Emphasis on Self-Report", in A.H. Tuma and J. Mason (eds.), *Anxiety and the Anxiety Disorders*, Erlbaum, Hillsdale, N.J.

Tennant, R., L. Hiller, R. Fishwick, S. Platt, S. Joseph, S. Weich, J. Parkinson, J. Secker and S. Stewart-Brown (2007), "The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation", *Health and Quality of Life Outcomes*, 5:63.

Thomas, D.L. and E. Diener (1990), "Memory Accuracy in the Recall of Emotions", *Journal of Personality and Social Psychology*, Vol. 59, No. 2, pp. 291-297.

Thompson, E.R. (2007), "Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS)", *Journal of Cross-Cultural Psychology*, Vol. 38, No. 2, pp. 227-242.

Tourangeau, R., K.A. Rasinski and N. Bradburn (1991), "Measuring Happiness in Surveys: A Test of the Subtraction Hypothesis", *Public Opinion Quarterly*, Vol. 55, pp. 255-266.

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.

Veenhoven, R. (2008), "The International Scale Interval Study: Improving the Comparability of Responses to Survey Questions about Happiness", in V. Moller and D. Huschka (eds.), *Quality of Life and the Millennium Challenge: Advances in Quality-of-Life Studies, Theory and Research*, Social Indicators Research Series, Vol. 35, Springer, pp. 45-58.

Vittersø. J., R. Biswas-Diener and E. Diener (2005), "The Divergent Meanings of Life Satisfaction: Item Response Modeling of the Satisfaction With Life Scale in Greenland and Norway", *Social Indicators Research*, Vol. 74, pp. 327-348.

Wänke, M. and N. Schwartz (1997), "Reducing Question Order Effects: The Operation of Buffer Items", in L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo and N. Schwarz (eds.), *Survey Measurement and Process Quality*, Wiley, Chichester, pp. 115-140.

Wanous, J.P., A.E. Reichers and M.J. Hudy (1997), "Overall Job Satisfaction: How Good are Single-Item Measures?", *Journal of Applied Psychology*, Vol. 82, No. 2, pp. 247-252.

Warr, P., J. Barter and G. Brownbridge (1983), "On the Independence of Positive and Negative Affect", *Journal of Personality and Social Psychology*, Vol. 44, No. 3, pp. 644-651.

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.

Watson, D. and L.A. Clark (1992), "On Traits and Temperament: General and Specific Factors of Emotional Experience and their Relation to the Five-Factor Model", *Journal of Personality*, Vol. 60, No. 2, pp. 441-476.

Watson, D., L.A. Clark and A. Tellegen (1988), "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales", *Journal of Personality and Social Psychology*, Vol. 54, No. 6, pp. 1063-1070.

Watson, D. and A. Tellegen (2002), "Aggregation, Acquiescence, and the Assessment of Trait Affectivity", *Journal of Research in Personality*, Vol. 36, pp. 589-597.

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.

Winkelmann, L. and R. Winkelmann (1998), "Why are the unemployed so unhappy? Evidence from panel data", *Economica*, Vol. 65, pp. 1-15.

Winkielman, P., B. Knäuper and N. Schwartz (1998), "Looking Back in Anger: Reference Periods Change the Interpretation of Emotion Frequency Questions", *Journal of Personality and Social Psychology*, Vol. 75, No. 3, pp. 719-728.

Wright, D.B., G.D. Gaskell and C.A. O'Muircheartaigh (1994), "How Much is 'Quite a bit'? Mapping Between Numerical Values and Vague Quantifiers", *Applied Cognitive Psychology*, Vol. 8, pp. 479-496.

Yardley, J.K. and R.W. Rice (1991), "The Relationship Between Mood and Subjective Well-Being, *Social Indicators Research*, Vol. 24, No. 1, pp. 101-111.