OECD Guidelines on Measuring Subjective Well-being

OECD

OECD Better Life Initiative

OECD Guidelines on Measuring Subjective Well-being

BETTER POLICIES FOR BETTER LIVES

Please cite this publication as:
OECD (2013), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing.
http://dx.doi.org/10.1787/9789264191655-en

ISBN 978-92-64-19164-8 (print)
ISBN 978-92-64-19165-5 (PDF)

European Union
Catalogue number: KE-31-13-501-EN-C (print)
Catalogue number: KE-31-13-501-EN-N (PDF)
ISBN 978-92-79-28316-1 (print)
ISBN 978-92-79-28315-4 (PDF)

Foreword

Understanding and improving well-being requires a sound evidence base that can inform policy-makers and citizens alike where, when, and for whom life is getting better. These Guidelines have been produced under the OECD's Better Life Initiative *– a pioneering project launched in 2011, which aims to measure society's progress across eleven domains of well-being, ranging from income, jobs, health, skills and housing, through to civic engagement and the environment. Subjective well-being – i.e. how people think about and experience their lives – is an important component of this overall framework. To be most useful to governments and other decision-makers, however, subjective well-being data need to be collected with large and representative samples and in a consistent way across different population groups and over time.*

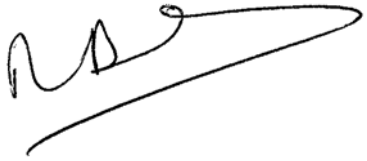
These Guidelines mark an important turning point in our knowledge of how subjective well-being can, and should, be measured. Not long ago, the received wisdom was that "we don't know enough" about subjective well-being to build it into measures of societal progress. However, as the evidence documented in these Guidelines shows, we in fact know a lot – perhaps more than we realised until we gathered all the relevant material for this report – and in particular that measures of subjective well-being are capable of capturing valid and meaningful information.

However, like all self-reported measures, survey-based measures of subjective well-being, are sensitive to measurement methodology. A large part of this report is therefore devoted to explaining some of the key measurement issues that both data producers and users need to know about. Comparable data require comparable methods, and a degree of standardisation that will require both determination and co-operation to achieve.

Subjective well-being data can provide an important complement to other indicators already used for monitoring and benchmarking countries performance, for guiding people's choices, and for designing and delivering policies. Measures of subjective well-being show meaningful associations with a range of life circumstances, including the other dimensions of well-being explored in the Better Life Initiative. *However, because a variety of factors affect how people experience and report on their lives, including factors such as psychological resilience in the face of adversity, and potential cultural and linguistic influences that are not currently well-understood, subjective well-being can only tell part of a person's story. These data must therefore be examined alongside information about more objective aspects of well-being, to provide a full and rounded picture of how life is.*

As for any new area of statistics, there is still much to be learned. These guidelines set out what we currently know about good practice. Research on both the measurement and the determinants of subjective well-being is rapidly advancing. As our knowledge grows, good practice will need to be updated. Nonetheless, it is important to recognise just how far we have come in recent years. These

Guidelines represent the first attempt to provide international recommendations on data collection, including some prototype question modules. Although this report is more of a beginning than an end, I believe it represents an important step forward.



Martine Durand
OECD Chief Statistician
Director of the OECD Statistics Directorate

Acknowledgments

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

These guidelines were produced as part of the work programme of the OECD Statistics Committee, whose delegates have reviewed the report. The report was prepared by Conal Smith and Carrie Exton. An expert advisory group (Tracey Chester, Ed Diener, David Halpern, John Helliwell, Alan Krueger, Bartek Lessaer, Chris Mackie, Robert Manchin, Ken Prewitt, Sue Taylor, and Takahashi Yoshiaki) provided valuable comments and advice on the drafting of the report. Additional comment and advice was received from a range of other experts in the field, including Saamah Abdullah, Bob Cummins, Angus Deaton, Paul Dolan, Stephen Hicks, Felicia Huppert, Susumu Kuwahara, Richard Layard, Filomena Maggino, Martin Ravallion, Layla Richroch, Arthur Stone, Richard Suzman, and Ruut Veenhoven. The report has benefited from the contributions and comments from Marco Mira d'Ercole and Martine Durand (OECD Statistics Directorate) as well as staff from other OECD directorates, and the national statistical agencies of OECD member countries. Generous support from the National Institute of Ageing was instrumental in the production of this report.

Table of contents

Tables

Figures

Overview and recommendations

What are these guidelines?

These guidelines provide advice on the collection and use of measures of subjective well-being. They are intended to provide support for national statistical offices and other producers of subjective well-being data in designing, collecting, and publishing measures of subjective well-being. In addition, the guidelines are designed to be of value to users of information on subjective well-being.

The guidelines provide information on the validity of subjective well-being measures; discuss the main methodological issues in developing questions to collect information on subjective well-being; present best practice in the measurement of subjective well-being; and provide guidance on the analysis and reporting of subjective well-being measures. A number of prototype question modules relating to different aspects of subjective well-being are also included.

These guidelines should be viewed as providing advice on best practice rather than being a formal statistical standard. At present, countries differ in terms of how much interest they have in information on subjective well-being, and in terms of the ability of national statistical offices to collect such data. The role of the guidelines, therefore, is primarily to assist data producers in meeting the needs of users by bringing together what is currently known on how to produce high quality, comparable measures of subjective well-being. As an international governmental organisation, the OECD has a particular interest in encouraging international comparability of data, and this is one of the key objectives of this report.

These guidelines aim to contribute to greater consistency in measurement of subjective well-being in official statistics. In particular, the guidelines are intended to:

- Improve the quality of subjective well-being measures collected by national statistical offices, by providing best practice in terms of question wording and survey design.
- Improve the usefulness of the data collected by setting out guidelines on the appropriate frequency, survey vehicles, and co-variates when collecting subjective well-being data.
- Improve cross-country comparability of subjective well-being measures by establishing common concepts, classifications, and methods that national statistical agencies could use.
- Provide advice and assistance to data users when analysing subjective well-being data.

What is subjective well-being?

The measurement of subjective well-being is often assumed to be restricted to measuring "happiness". In fact, subjective well-being covers a wider range of concepts than just happiness. For the purposes of these guidelines, a relatively broad definition of subjective well-being is used. In particular, subjective well-being is taken to be:[1]

> *Good mental states, including all of the various evaluations, positive and negative, that people make of their lives and the affective reactions of people to their experiences.*

This definition is intended to be inclusive, encompassing the full range of different aspects of subjective well-being commonly identified by reaserch in this field. It includes first and foremost measures of how people experience and evaluate their life as a whole. However, the guidelines also provide advice on measuring people's experience and evaluations of particular domains of life, such as satisfaction with their financial status or satisfaction with their health status, as well as measures of "meaningfulness" or "purpose" in life (often described as "eudaimonic" aspects of subjective well-being). This definition of subjective well-being hence encompasses three elements:

- *Life evaluation* – a reflective assessment on a person's life or some specific aspect of it.
- *Affect* – a person's feelings or emotional states, typically measured with reference to a particular point in time.
- *Eudaimonia* – a sense of meaning and purpose in life, or good psychological functioning.

The guidelines do not address subjective measures of objective concepts, such as self-rated health or perceived air quality. While the measurement tools for questions of this sort are "subjective", the subject matter being investigated is not, i.e. it can be observed by a third party. Some advice is provided, however, on measuring people's evaluations of specific domains of life, such as their *satisfaction with* their financial status or their health status. What is specific about the concept of subjective well-being as presented in this report, is that only the person under investigation can provide information on their evaluations, emotions and psychological functioning – it is people's own views that are the subject of interest.

Why have these guidelines been produced?

Notions of subjective well-being or happiness have a long tradition as central elements of quality of life. However, until recently, these concepts were often deemed beyond the scope of quantitative measurement. In the past two decades, however, an increasing body of evidence has shown that subjective well-being can be measured in surveys, that such measures are valid and reliable, and that they can inform policy making. This evidence has been reflected in the exponential growth of research in this field.[2]

Reflecting the increasing interest in subjective well-being from both researchers and policy-makers, the *Report of the Commission on the Measurement of Economic Performance and Social Progress* (Stiglitz et al., 2009) recommended that national statistical agencies collect and publish measures of subjective well-being. In particular, the Commission noted that:

> *Recent research has shown that it is possible to collect meaningful and reliable data on subjective well-being. Subjective well-being encompasses three different aspects: cognitive evaluations of one's life, positive emotions (joy, pride), and negative ones (pain, anger, worry). While these aspects of subjective well-being have different determinants, in all cases these determinants go well beyond people's income and material conditions… All these aspects of subjective well-being should be measured separately to derive a more comprehensive measure of people's quality of life and to allow*

a better understanding of its determinants (including people's objective conditions). National statistical agencies should incorporate questions on subjective well-being in their standard surveys to capture people's life evaluations, hedonic experiences and life priorities (p. 216).

Following on from the *Commission on the Measurement of Economic Performance and Social Progress*, an increasing number of statistical agencies have launched initiatives aimed at measuring subjective well-being.

While subjective well-being has been examined extensively in the academic literature, including some consideration of which subjective well-being measures to collect, and how to collect them, no consistent set of guidelines for national statistical agencies drawing on this research currently exist. For official measures of subjective well-being to be useful, these official measures should be collected in a consistent manner, which, in turn, requires a consensus on the best methodology to adopt. This is the main motivation for developing commonly accepted guidelines around the measurement of subjective well-being that draw on the best evidence available so far. These guidelines will need to be revised in the future as more information becomes available on subjective well-being.

How are the guidelines intended to be used?

The guidelines are intended to provide both a resource for data producers developing their own surveys, as well as a guide for ensuring that the data collected will be more internationally comparable. For users of the guidelines who are interested in developing their own questions, the guidelines could be used as a reference book. For users for whom international comparability is a priority, the guidelines also include more specific proposals on good practice, as it currently stands.

Chapter 1 provides an overview of the concepts being measured and assesses what is known about the validity and reliability of different types of measure. This is aimed to inform decisions about what aspects of subjective well-being should be measured and the degree to which such measures should be treated as experimental. Chapter 2 focuses on methodological issues in question and survey design and implementation. This part of the guidelines is intended primarily to support survey designers in developing questions on subjective well-being and to identify strategies for minimising bias due to different measurement effects – including not just how questions are worded, but also how surveys are implemented. Chapter 3 describes good practice in the measurement of subjective well-being. In particular, it includes practical recommendations on the issues of sample and survey design, building on evidence from the methodological issues discussed earlier. Chapter 4 provides advice on how best to report measures of subjective well-being as well as guidance for analysing this type of data.

An important part of the guidelines are the prototype question modules, introduced in Chapter 3, and attached at the end of the document, in Annex B (A through to F). In recognition of the different user needs and resources available to statistical producers, the guidelines do not present a single approach to gathering information on subjective well-being. Instead, six question modules are provided. Each question module focuses on a distinct aspect of subjective well-being.

Question Module A contains the core measures for which international comparability is the highest priority. These are the measures for which there is the most evidence for their validity and relevance, where results are best understood, and where policy uses are the most developed. Module A covers both life evaluation and affect measures, and all

national statistical agencies are encouraged to implement it in its entirety. A single experimental eudaimonic measure is also included in this core module. When it is not possible to collect the full core module, the primary life evaluation measure outlined in the module should be used at the minimum.

Modules B through to E are focused on specific aspects of subjective well-being. These modules are not necessarily intended to be used in their entirety or unaltered, but provide a resource for national statistical agencies developing their own questionnaires.

The six modules are listed below. For those highlighted as *recommended*, national statistical offices are encouraged to implement them in their entirety; the other modules are intended as a *resource* for data producers developing more detailed questionnaires.

Recommended for household surveys:

- A. Core measures.

Resource for household surveys:

- B. Life evaluation.
- C. Affect.
- D. Eudaimonic well-being.
- E. Domain evaluation.

Recommended for time use surveys:

- F. Experienced well-being.

Conclusions and recommendations

The guidelines include many recommendations. A summary from each of the chapters is presented here. Where more detail is needed, or where there is interest in why the recommendations take the form that they do, readers can refer to the main report. The overview is organised in four sections, mirroring the structure of the main report. Each section provides a short summary of the contents of the corresponding part of the report followed by the relevant recommendations.

1. Concept and validity

There is a large body of evidence on the reliability and validity of measures of subjective well-being and on the methodological challenges involved in collecting and analysing such data. Indeed, given the academic interest in the topic and the challenging nature of the subject, the body of evidence on the strengths and weaknesses of measures of subjective well-being may even exceed that available for many measures regularly collected as part of official statistics (Smith, 2013). While measures of subjective well-being have some important limitations, there is no strong case for simply considering them as falling "beyond the scope" of official statistics. Although subject to some methodological limitations, it is clear that for many potential uses, measures of subjective well-being, when carefully collected, are able to meet the basic standard of "fitness for purpose".

However, there are also areas where measures of subjective well-being are more strongly affected by how the measure is collected and by potentially irrelevant characteristics of the respondent than is the case for other official statistics. This does not imply that measures of subjective well-being should be rejected outright, but highlights two important points. First, official measures of subjective well-being should be collected – possibly as experimental data

series – in order to provide the basis to resolve some of the outstanding methodological issues. Second, information on the nature of the most significant methodological issues should be available to producers of subjective well-being data, and a common approach to dealing with these issues should be developed.

The main points with respect to the quality of subjective well-being measures are summarised under the headings relevance, reliability, and validity.

Relevance. Measures of subjective well-being have a wide variety of potential uses and audiences. These can be classified under four main headings:

- Complement other outcome measures.
- Help better understand the drivers of subjective well-being.
- Support policy evaluation and cost-benefit analysis, particularly where non-market outcomes are involved.
- Help in identifying potential policy problems.

Reliability. Reliability concerns the extent to which a measure yields consistent results (i.e. whether it has a high signal-to-noise ratio):

- Test-retest scores for measures of subjective well-being are generally lower than is the case for commonly collected statistics such as education and income, but higher than those found for more cognitively challenging economic concepts (such as household expenditure).
- The more reliable multi-item measures of subjective well-being, such as the satisfaction with life scale, exhibit higher reliability, although still less than for demographic or educational statistics.
- Looking at country averages, the reliability of life satisfaction measures is generally well above the required threshold for acceptable reliability.
- Measures of affect have, as expected, lower reliability than is the case for evaluative measures because moods change more frequently.
- There is relatively little evidence on the reliability of eudaimonic measures.

Validity. Validity is the extent to which an indicator actually captures the underlying concept that it purports to measure:

- Evidence strongly suggests that measures of both life evaluation and affect capture valid information.
- The evidence base for eudaimonic measures is less clear. While some specific measures – such as those relating to "meaning" and "purpose" clearly capture unique and meaningful information, the picture with respect to eudaimonia as a whole is more ambiguous. This suggests that further work is needed before a definitive position can be taken on the validity of these measures.
- While a range of issues could place limits on the validity of subjective measures of well-being, many of these have either a marginal impact on fitness for purpose (i.e. they do not substantively affect the conclusions reached) or can be dealt with through appropriate survey design and carefully considered analysis and interpretation of the data.
- Despite evidence that cultural factors do not substantively bias multi-variate analysis, there are reasons to be cautious about cross-country comparisons of levels of subjective well-being.

2. Methodological considerations in the measurement of subjective well-being

Much like other survey-based measures, 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 – whether comparisons are to be made over time or between groups of respondents, the guidelines argue in favour of adopting a consistent measurement approach across all survey instruments, study waves and countries wherever possible, to limit the additional variance potentially introduced by differing methodologies. Of course, the exception to this is where experimental statistics are being collected for the explicit purpose of examining methodological factors. In this case, it is important to vary single elements of survey design, one-by-one, in a systematic manner.

Recommendations: Question wording and response formats

- In terms of survey design, question 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 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. While evaluative and eudaimonic measures are intended to capture constructs spanning a longer time period, there is less evidence 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 using a 0-10 point numerical scale, anchored by verbal labels which represent conceptual absolutes (such as *completely satisfied/completely dissatisfied*). On balance, it is preferable to label scale interval-points (between the anchors) with numerical, rather than with verbal, labels.
- The order in which response categories are presented to respondents may be particularly important in 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 the lowest (e.g. 0) to the highest (e.g. 10) may help reduce respondent burden.
- In the case of affect measures, unipolar scales (i.e. 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. This is because, for conceptual reasons, it is helpful measure positive and negative affect separately, rather than combining them in a single bipolar (*very sad/very happy*) question. For life evaluations and eudaimonia, there is less evidence on scale polarity. The available information suggests that bipolar and unipolar measure produce very *similar results for life* evaluation measures, although bipolar scales may be confusing for respondents when evaluative questions are negatively-framed.

Recommendations: Question order and context effects

- Question order effects can be a significant problem, but one that can be managed by asking subjective well-being questions before other sensitive survey items. Where this is not possible, the use of introductory text or of other questions can buffer the impact of context.

- Order effects also exist *within* sets of subjective well-being questions. Question modules should include only one primary evaluative measure, flow from the general to the specific, and use consistent ordering of positive and negative affect (to reduce the risk that asking negative questions first may bias subsequent responses to positive questions, and vice versa).

Recommendations: Survey mode and timing

- The use of different survey modes can produce differences in subjective well-being data – although the significance and magnitude of difference varies considerably across studies. Given the trade-offs to be considered when selecting between survey modes, there is no clear "winner" – although from a data quality perspective, face-to-face interviewing has a number of advantages.
- Where mixed-mode surveys are used, it is 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. This is especially important where sample characteristics may influence the mode of survey response (e.g. regional or demographic variations).
- Aspects of the wider survey context, such as the day of the week in which the survey is conducted and day-to-day events occurring around the time of the survey, can influence affective measures, but this should not be regarded as error. There is also some evidence that rare and/or significant events can impact on life evaluations. It is critical, therefore, 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 that are sampled.

Recommendations: Response styles and international comparability

- Response styles present particular challenges for data interpretation when they vary systematically between countries, or between population sub-groups within countries. However, 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 current approach for guarding against biases introduced by response styles is to adopt sound survey design principles that minimise the risk that respondents 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.

Conclusion: Methodological considerations

Perhaps *because* of concerns about their use, quite a lot *is* known about how subjective well-being measures behave under different measurement conditions. 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, questions remain regarding the "optimal" way to measure subjective well-being. National statistical agencies are in a unique position to improve the evidence base, by providing the data that can answer those questions for which large and more nationally-representative samples are required.

3. Measuring subjective well-being

This section presents best practice in measuring subjective well-being. It covers both the range of concepts to be measured and the best approaches for measuring them. This includes considering issues of sample design, survey design, data processing and coding and questionnaire design.

Survey vehicles

Subjective well-being measures are relevant in a wide range of contexts. Of particular importance for monitoring progress is the inclusion of such measures in integrated household surveys and general social surveys. Time use surveys are the key vehicle for collecting detailed information on affect and its antecedents but it is possible to collect useful information on affect from other household surveys.

Measures of subjective well-being are also relevant to victimisation surveys, health surveys, and special topic surveys. In particular, special topic surveys are excellent vehicles for exploring aspects of subjective well-being in more depth, although they cannot be used to monitor changes in well-being over time due to their "one-off" nature.

Including measures of subjective well-being in panel surveys is important for research into causality and the drivers of subjective well-being.

What other information should be collected: Co-variates and analytical and variables

The precise range of co-variates to collect alongside measures of subjective well-being will vary with the specific aspect of subjective well-being that is of interest and with the research question being examined. Despite this, it is possible to present some general guidelines on the most important information that should be collected alongside measures of subjective well-being:

- *Demographics:* Age, gender, marital status (legal marital status and social marital status), family type, number of children, household size, and geographic information.
- *Material conditions:* Household income, consumption, deprivation, housing quality.
- *Quality of life:* Employment status, health status, work/life balance, education and skills, social connections, civic engagement and governance, environmental quality, personal security.
- *Psychological measures:* aspirations and expectations about the future, which form part of the frame of reference[3] that individuals use when evaluating their lives or reporting their feelings.

Time use diaries. Although all of the measures identified as relevant to household surveys are equally relevant to time use surveys, the use of time diaries allows the collection of information on additional co-variates in a way that is not possible in standard household surveys (e.g. activity classification, with whom an activity was performed, location where the activity took place). This is particularly useful to have where information on aspects of subjective well-being, such as affect, are collected in the diary itself.

Target population

The target age group for measures of subjective well-being will vary with respect to the goals of the research programme. For example, in the context of research on retirement income policies, it may be appropriate to limit the target population to persons aged 65 or

older. In general, however, measures of subjective well-being would usually be collected for all the adult population (aged 15 years and older):

- In all cases, the sampling frame must produce a representative sample of individuals or households as if all individuals are personally interviewed.
- Evidence suggests that children are capable of answering questions from age 11 with respect to both measures of life evaluation and affective state.
- Proxy responses, which might be appropriate for some types of data (income, marital status, age) are not valid with respect to subjective well-being.

Frequency and duration of enumeration

For the most important core measures used to monitor well-being, an annual time series should be regarded as the essential minimum in terms of frequency of enumeration:

- Ideally, enumeration would take place over a full year, and would include all days of the week including holidays.
- Where a year-long enumeration period is not possible, enumeration should, as far as possible, be spread proportionately over all the days of the week.

Sample size

Large samples are desirable for subjective well-being as for any topics, as they reduce the standard error of estimates and allow a more precise estimate and a greater degree of freedom with respect to producing cross-tabulations and analysis for population sub-groups.

Mode

In terms of data quality, computer assisted personal interviewing (CAPI) with show cards is currently considered the ideal choice for collecting subjective well-being data:

- Where other modes of interviewing – such as computer assisted telephone interviewing (CATI) or computer assisted self interview (CASI) – are used, it is important that data producers collect information to enable the impact of mode effects to be estimated.
- National statistical agencies, in particular, should test experimentally the impact of different survey modes on responses to the core measures of subjective well-being, and publish the results along with any results from CATI or CASI surveys.

Question placement

Question placement within a survey can have a considerable impact on responses. A number of actions can minimise this effect:

- *Place important subjective well-being questions near the start of the survey.* Although placing questions early in a survey does not eliminate all of the problems associated with context effects, it is the best strategy available and should be pursued where possible. In particular, for the core measures of subjective well-being, for which international or time series comparisons are critical, these questions should be placed directly after the initial screening questions that lead to a respondent's inclusion in the survey. The core measures module included in Annex B is intended to be placed at the start of a survey.
- *Avoid placing the subjective well-being questions immediately after questions likely to elicit a strong emotional response or that respondents might use as a heuristic for determining their response to the subjective well-being question.* This would include questions on income,

social contact, labour force status, victimisation, political beliefs, or any questions suggesting social ranking. The best questions to precede subjective questions are relatively neutral factual demographic questions.

- *Make use of transition questions to refocus respondent attention*. One technique that has been used to address contextual effects resulting from a preceding question is using a transition question designed to focus the respondent's attention on their personal life. However, transition questions can also introduce their own context effects. For example, drawing attention to a respondent's personal life may lead them to focus on personal relationships or family when answering subsequent questions about life overall. Development of effective transition questions is a priority for future work.
- *Use of introductory text to distinguish between question topics*. Well worded text preceding each question or topic can serve as a buffer between measures of subjective well-being and sensitive questions. However, there is little evidence on the effectiveness or optimal phrasing of such introductory text. A standard introductory text has been included in each of the prototype question modules included in Annex B to this document. Further cognitive testing or experimental analysis of the impact of different types of introductory text would be highly valuable.

Choice of questions

It is recommended that any changes to existing questions are phased-in using parallel samples, so that the impact of the change can be fully documented and examined. This will provide insights into the impact of changes in methodology, and provide agencies with information for adjusting previous data sets.

Translation

Because question wording matters to how people respond to questions on subjective well-being, the translation of questions is of high importance. Potential issues arising from translation cannot be entirely eliminated, but they can be managed through an effective translation process. A robust translation process, including back translation, is therefore essential.

Interviewer training

To manage risks around respondent attitudes to questions on subjective well-being, it is imperative that interviewers are well-briefed, not just on what concepts the questions are trying to measure, but also on how the information collected will be used. This is essential for interviewers to build a good relation with respondents, and can improve respondents' compliance and data quality.

Measures of subjective well-being are relatively non-problematic for respondents to answer. Rates of refusal to respond are low, both for life evaluations and for measures of affect. However, statistical providers should consider how best to manage the risks associated with questions that are distressing to respondents. Although these risks should not be overstated – they apply mainly to eudaimonic questions and to a small proportion of respondents – such issues should be dealt with effectively.

4. Output and analysis of subjective well-being measures

It is difficult to provide a succinct list of recommendations relating to the analysis of subjective well-being data, as much of the advice is contingent on the goals of the analysis. Nonetheless, several recommendations that can be made with respect to publishing measures of subjective well-being:

- Using subjective well-being data to complement other measures of well-being requires producers of statistical information to regularly collect and release high quality nationwide data from large and representative samples.
- It will be important, especially when reporting the results of national surveys, to provide a full description of the indicators used – including the underlying constructs of interest, and what they might reflect in addition to "happiness".
- Data releases should include information on both the level and distribution of measures of subjective well-being.

Reporting central tendency and level of subjective well-being measures

Summary statistics of central tendency such as mean, median, and mode provide a useful way of presenting and comparing the level of subjective well-being in a population in a single number. Information on levels may also be presented as the proportion of the relevant population above or below a particular threshold. Threshold based measures can be a useful way to summarise findings in a simple manner for non-expert audiences. However, the choice of threshold is a critical consideration, as threshold-based reporting can produce distortions and mask important changes in the data:

- The mean is useful as a summary statistic of the level of subjective well-being.
- Where possible, both the mean score and the proportion of the population above or below specific thresholds should be reported in data releases.
- When publishing data based on a threshold, it is important to check that the change in the proportion of the population above or below the threshold is consistent with the picture that emerges from a look at changes in the distribution as a whole. The choice of threshold should also be explained.

Reporting the distribution of subjective well-being measures

Given the limited number of response options associated with measures of subjective well-being (typically 0 to 10 for most of the questions proposed in the guidelines), it may be possible to publish data showing the entire distribution of responses:

- Where summary measures of dispersion are required, the inter-quartile range (i.e. the difference between individuals at the 25th percentile and individuals at the 75th percentile of the distribution), or the point difference between quantiles (e.g. the 90th and the 10th percentile) may be preferred.

Analysis of subjective well-being data

It is difficult to provide a succinct list of recommendations relating to the analysis of subjective well-being data, as much of the advice is contingent on the goals of the analysis. However, this section of the guidelines covers topics such as:

- Issues in the analysis and interpretation of descriptive statistics on subjective well-being. This includes the interpretation of change over time, differences between groups, and the risk of cultural "bias" in cross-country comparisons.
- Analyses of the drivers of well-being, including managing issues such as shared method variance, omitted variables, reverse and two-way causality, frame of reference effects and hedonic adaptation.
- How subjective well-being data might be used to inform the appraisal, design and evaluation of policy options, including the potential use of subjective well-being data in cost-benefit analysis.

What is next?

These guidelines do not aim to provide the "final word' on the measurement of subjective well-being. Although some aspects of the measurement of subjective well-being – such as questions on overall satisfaction with life – are well understood, other potentially important measures currently draw on much weaker evidence bases. It is expected that the evidence base on subjective well-being will develop rapidly over the next few years. In particular, to the extent that national statistical offices start regularly collecting and publishing data on subjective well-being, many methodological questions are likely to be resolved as better data becomes available, and an increasing body of knowledge will accumulate on the policy uses of subjective well-being data.

It is envisaged that these guidelines will be followed up by a review of progress on the measurement of subjective well-being over the next few years, with a view to deciding whether the guidelines need revising and whether it is possible and desirable to move towards a greater degree of international standardisation. The intent is that this review will build on information collected by national statistical agencies, and will consider the feasibility of moving towards a more formal international standard for the measurement of subjective well-being.

Notes

1. The definition used here draws largely on Diener et al. (2006).
2. During the 1990s there was an average of less than five articles on happiness or related subjects each year in the journals covered by the *Econlit* database. By 2008 this had risen to over fifty each year.
3. "Frame of reference" refers to the situation or group on which respondent's base comparisons when formulating a judgement about their lives or feelings. The respondent's knowledge of how others live and their own prior experiences can influence the basis on which judgements are reached about the respondent's current status.

Introduction

Notions of subjective well-being or happiness have a long tradition as central elements of quality of life, but until very recently these concepts were generally deemed beyond the scope of statistical measurement. Over the last two decades, however, an increasing body of evidence has shown that subjective well-being can be measured in surveys, that such measures are valid and reliable, and that they can usefully inform policy-making. This evidence has been reflected in an exponential growth in the economic literature on measures of subjective well-being.[1]

Reflecting the increasing interest in subjective well-being from both researchers and policy-makers, the *Report by the Commission on the Measurement of Economic Performance and Social Progress* (2009) recommended that national statistical agencies collect and publish measures of subjective well-being. In particular, the Commission noted that:

> *Recent research has shown that it is possible to collect meaningful and reliable data on subjective well-being. Subjective well-being encompasses three different aspects: cognitive evaluations of one's life, positive emotions (joy, pride), and negative ones (pain, anger, worry). While these aspects of subjective well-being have different determinants, in all cases these determinants go well beyond people's income and material conditions… All these aspects of subjective well-being should be measured separately to derive a more comprehensive measure of people's quality of life and to allow a better understanding of its determinants (including people's objective conditions). National statistical agencies should incorporate questions on subjective well-being in their standard surveys to capture people's life evaluations, hedonic experiences and life priorities.*[2]

The guidelines presented here represent a step towards making the Commission's recommendations a reality. They are intended to provide guidance and assistance to data producers, and particularly national statistical agencies, in collecting and reporting measures of subjective well-being, as well as providing advice and assistance in the analysis of subjective well-being data to users of the data.

Motivation

Recent initiatives

The OECD recently characterised its mission as "better policies for better lives". This implies a concern with the nature and drivers of people's well-being. In order to develop better policies, it is essential to understand what constitutes "better lives" for the citizens of OECD countries. This concern with what constitutes well-being and how well-being should be measured has been reflected in OECD work, including the activities related to the OECD-hosted Global Project on Measuring the Progress of Societies and the associated series of World Forums on Statistics, Knowledge, and Policy held in Palermo (2004), Istanbul (2007), Busan (2009) and Delhi (2012). More recently, building on the foundations set out by the

Report of the Commission on the Measurement of Economic Performance and Social Progress (the Sen/Stiglitz/Fitoussi Commission), the OECD has developed tools that allow users to build their own measure of average well-being across countries, through the *Your Better Life Index*.

Following on from the *Commission on the Measurement of Economic Performance and Social Progress*, a number of statistical agencies have launched initiatives aimed at measuring subjective well-being. These include the UK initiative (launched in November 2010) to develop a new set of measures of national well-being (combining both subjective and objective measures) and the steps taken by Eurostat to develop a module on well-being for the 2013 wave of EU-SILC.[3] Similarly, the French national statistical office, INSEE, has developed a well-being module for the national component of EU-SILC and has collected information on affect in the *Enquête Emploi du temps 2009-2010*. In the United States, a well-being module has also been included in the most recent wave of the American Time Use Survey. Also, the US National Academy of Sciences has established a panel on *Measuring Subjective Well-Being in a Policy Relevant Framework*. In Italy, the national statistical office has recently published its first official measures of life satisfaction as part of its general social survey (*Indagine Multiscopo*). In the Netherlands, the national statistical office is currently scoping a module on subjective well-being for one of its surveys to go into the field (if approved) in late 2011/12. Plans to collect data on subjective well-being as part of their official statistical systems were also recently announced by Japan and Korea.

A number of national statistical agencies have collected data on subjective well-being even before the recommendations of the *Commission on the Measurement of Economic Performance and Social Progress*. In Canada, Statistics Canada has collected information on subjective well-being in the General Social Survey since 1985 and published this information as part of data releases from the survey for some time. The national statistical office of New Zealand also collects data on life satisfaction through the New Zealand General Social Survey, and this forms a core component of its data release. The Australian Bureau of Statistics has collected information on subjective well-being in a number of vehicles, including the 2001 National Health Survey and the Australian General Social Survey.

The need for guidelines

The use of international concepts and measurement methodology is fundamental to official statistics. Such standards contribute to quality by ensuring that best practice is followed internationally and that official statistics are internationally comparable. This is reflected in the *UN Fundamental Principles of Official Statistics*. In particular, principle 9 states that:

> *The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.*

Although measures of subjective well-being are now available from both official and non-official sources for an increasing range of countries, these measures currently lack commonly accepted international guidelines for their collection and dissemination. The only measures of subjective well-being on which cross-country comparisons can currently be made on a consistent basis are derived from non-official sources, and these face limitations associated with relatively small sample sizes, limitations of sample design and low survey response rates (Box 1).

Box 1. **Non-official sources of subjective well-being data**

Measures of subjective well-being are not currently collected in a systematic and consistent way across OECD national statistical agencies. While a number of OECD countries do collect measures of subjective well-being as part of their official statistics, and in some cases have been doing so for some time, official measures currently lack the consistency needed for them to be used as the basis for international comparisons. There are, however, a number of datasets currently available that contain measures of subjective well-being covering a wide range of countries. Indeed, much of the current body of knowledge regarding the validity and properties of measures of subjective well-being is derived from the analysis of these non-official datasets.

The two largest datasets containing comparable measures of subjective well-being are the Gallup World Poll and the World Values Survey. The Gallup World Poll started in 2005, and now covers 132 countries; the sample size is about 1 000 respondents per country per wave, with plans to increase such sample size to 4 000 respondents in all countries with a population of 25 million and over by 2012. The Gallup World Poll is an annual survey and includes measures of life evaluation and a range of questions related to current mood and emotional experiences (affect). The World Values Survey has a longer history (although with uneven sampling quality), with the first wave having been collected between 1981 and 1983 and covering 15 countries. There have been four subsequent waves, with the most recent wave collected between 2005 and 2008 and covering 56 countries. A sixth wave is currently being collected (2011-12). The World Values Survey contains measures of life evaluation and overall happiness, as well as more focused measures of experienced mood and aspects of psychological well-being in the more recent waves.

While the Gallup World Poll and the World Values Survey are usually taken as providing a reference point for questions on subjective well-being across countries, there are a number of additional surveys that complement these in various ways. The European Social Survey provides information on a number of aspects of subjective well-being for a varying range of European countries between 2002 and 2010. In the 2006 wave of the European Social Survey, a module was included to collect detailed information on the "eudaimonic" aspects of well-being (i.e. meaning, purpose, flourishing), thus expanding the range of subjective well-being concepts measured beyond evaluations and affect. A repeat of this module will be carried out in 2012. In addition, the triennial European Quality of Life Survey contains extensive information on subjective well-being.

Eurobarometer is a regular opinion survey covering European Union nations that has been collected since 1973. Although the subjective well-being questions contained in Eurobarometer are relatively limited, they provide the longest unbroken time series for measures of subjective well-being for a cross-section of countries. Similar question have also been included in several waves of the Latinobarómetro.

In addition to these cross-sectional surveys, a number of panel surveys have been widely used by researchers to analyse subjective well-being. In particular, the German Socio-Economic Panel and the British Household Panel Study are high-quality panel surveys that include information on subjective well-being. The German Socio-Economic Panel dates back to 1984 and runs to the present day, with a total sample of over 12 000 households. By comparison, the British Household Panel Study dates back only to 1991, but has recently been integrated into the UK Household Longitudinal Study (also known as "Understanding Society"), with a total sample of over 40 000 households. Because both of these studies follow the same person through time, they have been crucial in allowing researchers to understand the interaction between unobserved personality traits, life-events, environmental changes (including policy changes) and responses to subjective well-being questions.

Although non-official data sources have provided much information on subjective well-being, they do have several distinct limitations. With the exception of the large panel studies, most non-official data sources have relatively small sample sizes that limit the conclusions which can be reached about changes in levels of subjective well-being and differences between groups. Many of the main non-official surveys also are affected by low response rates and have sample frames that are not as representative as is the case for official surveys. Finally, the developers of non-official surveys often have fewer resources available for cognitive testing and survey development than is the case for national statistical offices. Thus, although existing non-official data sources have provided a great deal of information on subjective well-being, there remain a range of questions that will not be answered until high-quality large-scale official surveys are available.

While the academic literature contains extensive information about which subjective well-being measures to collect and how to collect them, no consistent set of guidelines currently exist for national statistical agencies that wish to draw on this research. For official measures of subjective well-being to be useful as indicators of national progress, these official measures should be collected in a consistent manner, which, in turn, requires an agreed way to collect such measures. This drives the need for developing commonly accepted guidelines around the measurement of subjective well-being, even if such guidelines will need to be revised in the future as more information becomes available on subjective well-being.

Guidelines are also needed because subjective well-being measures are strongly affected by question structure and context, and the results from differently worded questions (or even a different ordering of similar questions) are likely to affect comparability. Yet comparability is a key point of interest for decision-makers, who will often want to benchmark the progress of one region, country or population group against another. While interpreting such comparisons can be difficult due to issues such as cultural biases in response styles, consistency in measurement can eliminate other potential sources of bias. It is important that, where there are differences in measured levels of subjective well-being, these are not falsely attributed some significance when, in fact, the difference actually reflects the impact of question wording or context.

The guidelines

Scope and objectives

The aim of the project is to prepare a set of guidelines addressed to national statistical offices on the collection and use of measures of subjective well-being. This includes first and foremost measures of how people experience and evaluate life as a whole. Over-arching measures of this sort have been the main focus for academic analysis of subjective well-being and are therefore the best understood measures of subjective well-being, including because they reflect people's experiences and evaluations of all the different aspects of life, and therefore bring the most additional information to existing outcome measures such as income, health, education and time use. Despite this, the guidelines do also attempt to provide advice on people's evaluations of particular domains of life, such as satisfaction with their financial status or satisfaction with their health status as well as "eudaimonic"[4] aspects of subjective well-being. These measures are both of high interest for policy purposes and also methodologically similar to the more general questions on overall subjective well-being.

The guidelines do not attempt to address subjective measures of objective concepts. Measures of this sort, such as self-rated health or perceived air quality, are outside the scope of this project. While the measurement technique for questions of this sort is subjective, the subject matter is not, and such questions pose different methodological issues in measurement.

This report will outline both why measures of subjective well-being are relevant for monitoring the well-being of people and for policy design and evaluation and why national statistical agencies have a critical role in enhancing the usefulness of existing measures. The report will identify the best approaches for measuring in a reliable and consistent way the various dimensions of subjective well-being and will provide guidance for reporting on such measures. The project also includes the development of prototype survey modules on subjective well-being that national and international agencies could take as a starting point when designing their national surveys and undertaking any further testing and development.

The production of a set of guidelines on measuring subjective well-being by the OECD is expected to contribute to greater consistency in measurement of subjective well-being in official statistics. In particular, these guidelines are intended to:

- Improve the quality of measures collected by national statistical offices by providing best practice in question wording and survey design.
- Improve the usefulness of data collected by setting out guidelines on the appropriate frequency, survey vehicles and co-variates when collecting subjective well-being data.
- Improve the international comparability of subjective well-being measures by establishing common concepts, classifications and methods that national statistical agencies could use.

These guidelines do not by any means represent the final word on the measurement of subjective well-being. Although some aspects of the measurement of subjective well-being – such as questions on overall satisfaction with life – are very well understood, other potentially important measures currently draw on much weaker evidence bases. It is to be expected that the evidence base on subjective well-being will develop rapidly over the next few years. In particular, to the degree that national statistical offices start regularly collecting and publishing data on subjective well-being, many methodological questions are likely to be resolved as better data becomes available, and an increasing body of knowledge will accumulate around the policy uses of subjective well-being data.

It is envisaged that these guidelines will be followed up by a review of progress on the measurement of subjective well-being over the next few years, with a view to deciding whether the guidelines need revising and whether it is possible and desirable to move towards a greater degree of international standardisation. The intent is that this review will build on information collected by national statistical agencies, and will consider the feasibility of eventual moves towards a more formal international standard for the measurement of subjective well-being.

The structure of the guidelines

The guidelines are organised in four chapters. Chapter 1 focuses on the issues of concept and validity. This chapter addresses the issue of what subjective well-being "is" and describes a conceptual framework for subjective well-being, including a clear over-arching definition of the scope of subjective well-being and how this relates to broader notions of quality of life. The issue of the reliability and validity of measures of subjective well-being is also addressed, with a review of the evidence on validity. Finally, the chapter provides an overview of some of the limitations of subjective well-being measures, setting out some of the known problems and shortcomings.

The second chapter focuses on the methodological issues that should inform the selection of measures of subjective well-being through surveys. This chapter is framed around issues of survey mode, survey flow and question design. In particular, the chapter covers the impact of issues related to question order, question placement within the survey, question wording, scale formats and labelling, day and time effects and biases due to social desirability. In addition to identifying the key methodological issues raised, Chapter 2 makes recommendations on the best approach to mitigate the effect of various sources of bias.

Chapter 3 sets out an over-arching strategy for the measurement of subjective well-being. This covers both the range of concepts that should be measured and the choice of survey vehicles for measuring them. Issues of sample design and the statistical units to be measured are discussed, as well as the most appropriate range of co-variates to collect along with the subjective measures of well-being. The specific suite of measures proposed will also be outlined.

The final chapter sets out guidelines for the output and analysis of subjective well-being data. The first section of the chapter coves the issues associated with basic reporting of subjective well-being data, including what constitutes meaningful change and issues such as whether to report average scores or the proportion of the population relative to a threshold of some sort. The second part of the chapter gives a more detailed treatment of the use of subjective well-being data.

Annex A of the guidelines provides illustrative examples of different types of subjective well-being questions that have been used previously throughout the world. It is intended primarily to help users of the guidelines understand the methodological references to specific question types made in Chapter 2. Annex B contains six prototype question modules for national statistical agencies and other producers of subjective well-being data to use as models for their own questions. Module A (core measures) contains a primary measure of subjective well-being that all data producers are strongly encouraged to include as a baseline measure, along with four additional questions that should also be regarded as highly desirable to collect wherever space permits. The remaining five modules provide more detailed information that may be used by data producers where more detailed information on one of the dimensions of subjective well-being is considered a priority.

Notes

1. During the 1990s there was an average of less than five articles on happiness or related subjects each year in the journals covered by the *Econlit* database. By 2008 this had risen to over fifty.
2. *Report by the Commission on the Measurement of Economic Performance and Social Progress,* Stiglitz, J.E., A. Sen and J.P. Fitoussi, 2009, p. 216.
3. The European Union Statistics on Income and Living Conditions (EU-SILC) is an instrument aimed at collecting timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions. It is run in all European Union countries and some outside the EU, including Turkey, Norway, Iceland, Croatia, Serbia and Switzerland.
4. The term "eudaimonic" derives from the Greek word *eudaimonia*, which Aristotle used to refer to the "good" life. Eudaimonia implies a broader range of concerns than just "happiness". While Aristotle argues that happiness is necessary for eudaimonia, he believes it is not sufficient. Modern conceptions of eudaimonic well-being, although differing from Aristotle in the detail, focus on subjective well-being perceived more broadly than simply one's evaluation of life or affective state.

Chapter 1

Concept and validity

The main focus of this chapter is to set the conceptual scope for the measurement of subjective well-being and to provide an overview of what is currently known about the statistical quality of subjective well-being measures. The chapter covers what is meant by subjective well-being, its relevance and why it should be measured, and reviews the evidence on the validity of different types of subjective well-being measure.

In the first section of the chapter a conceptual framework is proposed for measures of subjective well-being, which outlines the main elements of subjective well-being and how these relate to each other. This is necessary both because a clear view of what is being measured is logically prior to decisions about how to measure them and because subjective well-being covers a number of distinct concepts; it is therefore important to be clear about what exactly is covered by the guidelines.

The remainder of the chapter focuses on issues of statistical quality, particularly the *relevance* and *accuracy* of measures of subjective well-being. Relevance addresses the issue of why measures of subjective well-being are important and how they can be used. Accuracy, on the other hand, is concerned with the degree of validity of measures of subjective well-being. In particular, the chapter considers the validity of measures of subjective well-being, focusing on the notion of "fitness for purpose" with respect to specific user needs. There are a number of issues where the evidence on measures of subjective well-being is insufficient to form a clear view of fitness for purpose. For this reason, the final section of the chapter concludes by summarising known issues with data on subjective well-being, including problems with measures of subjective well-being and areas where further research is needed.

1. Conceptual framework

In measurement, it is important to be clear about the nature and scope of the concept being measured. This is particularly the case for a topic such as subjective well-being where the precise concept being measured is less immediately obvious than is the case for a more straight-forward concept such as income, consumption, age or gender. The validity of a statistical measure – as will be discussed in the following sections of this chapter – can be understood as the degree to which the statistical measure in question captures the underlying concept that it is intended to measure. A clear conceptual framework for subjective well-being is therefore essential before it is possible to discuss validity in any meaningful sense.

The first element of a conceptual framework for the measurement of subjective well-being is to define exactly what is meant by subjective well-being. This is important because there are potentially a wide range of subjective phenomena on which people could report, not all of which would necessarily fall under the heading of "well-being". It is also important to define subjective well-being in order to be able to communicate clearly what is being measured. Often, the measurement of subjective well-being is conflated with measuring "happiness"; however, this is both technically incorrect (there is more to

subjective well-being than happiness) and misleading, and thus lends support to sceptics who characterise the measurement of subjective well-being in general as little more than "happiology".[1]

Most experts characterise subjective well-being as covering a number of different aspects of a person's subjective state (Diener et al., 1999; Kahneman, Diener and Schwarz, 1999). However, there is room for some debate about exactly what elements should be included. For example, some analysts, such as Kahneman and Krueger (2006), focus primarily on the hedonic aspect of subjective experience, while others, such as Huppert et al. (2009), opt for a definition that includes measures of good psychological functioning as well as purpose in life. For the purposes of these guidelines, a relatively broad definition of subjective well-being is used. In particular, subjective well-being is taken to be:[2]

> *Good mental states, including all of the various evaluations, positive and negative, that people make of their lives, and the affective reactions of people to their experiences.*

This definition is intended to be inclusive in nature, encompassing the full range of different aspects of subjective well-being commonly identified. In particular, the reference to good mental functioning should be considered as including concepts such as interest, engagement and meaning, as well as satisfaction and affective states. Thus, in the terms of Diener (2006), "subjective well-being is an umbrella term for the different valuations people make regarding their lives, the events happening to them, their bodies and minds, and the circumstances in which they live". Such valuations are subjective, in that they are experienced internally (i.e. they are not assessments of some external phenomenon); they constitute aspects of well-being in that they relate to the pleasantness and desirability or otherwise of particular states and aspects of people's lives.

While the definition of subjective well-being used here is broad and potentially reflects the influence of a wide range of people's attributes and circumstances, it does not imply that subjective well-being is proposed as *the* single all-encompassing measure of people's well-being, with all other aspects having only instrumental value in achieving this. On the contrary, this definition is explicitly consistent with approaches that conceive of people's well-being as a collection of different aspects, each of them having intrinsic value. In measuring overall human well-being then, subjective well-being should be placed alongside measures of non-subjective outcomes, such as income, health, knowledge and skills, safety, environmental quality and social connections.

The definition of subjective well-being outlined above is relatively broad, and could give the impression that subjective well-being is a hopelessly vague concept. This is not the case. There is, in fact, general agreement among experts on the specific aspects that comprise subjective well-being (Dolan and White, 2007; Sen, Stiglitz and Fitoussi, 2009; ONS, 2011). In particular, a distinction is commonly made between life evaluations, which involve a cognitive evaluation of the respondent's life as a whole (or aspects of it), and measures of affect, which capture the feelings experienced by the respondent at a particular point in time (Diener, 1984; Kahneman et al., 1999). In addition to the distinction between evaluation and affect, a number of researchers argue that there is also a clear eudaimonic aspect of subjective well-being, reflecting people's sense of purpose and engagement (Huppert et al., 2009). The framework used here covers all three concepts of well-being:

- Life evaluation.
- Affect.
- Eudaimonia (psychological "flourishing").

Elements of subjective well-being

Life evaluation

Life evaluations capture a reflective assessment on a person's life or some specific aspect of it. This can be an assessment of "life as a whole" or something more focused. Such assessments are the result of a judgement by the individual rather than the description of an emotional state. Pavot and Diener et al. (1991) describe the process of making an evaluation of this sort as involving the individual constructing a "standard" that they perceive as appropriate for themselves, and then comparing the circumstances of their life to that standard. This provides a useful way to understand the concept of life evaluation, although in practice it is not clear whether the process of comparison is a conscious one if respondents more commonly use a heuristic to reach a decision.

There is evidence that the construct captured by life evaluation is closely related to that used by people when they make a conscious judgement that one course of action is preferable to another (Kahneman et al., 1999; Helliwell and Barrington-Leigh, 2010). The underlying concept being measured is thus, in some senses, relatively close to an economist's definition of utility. However, economists usually assume (at least implicitly) that the remembered utility on which people base their decisions is equivalent to the sum of momentary utilities associated with moment-by-moment experiences. This is not the case. Life evaluations are based on how people remember their experiences (Kahneman et al., 1999) and can differ significantly from how they actually experienced things at the time. In particular, the so-called "peak-end rule" states that a person's evaluation of an event is based largely on the most intense (peak) emotion experienced during the event and by the last (end) emotion experienced, rather than the average or integral of emotional experiences over time. It is for this reason that life evaluations are sometimes characterised as measures of "decision utility" in contrast to "experienced utility" (Kahneman and Krueger, 2006).[3] Despite this limitation, the fact that life evaluations capture the same sort of construct that people use when making conscious decisions and align closely to the conception of individual welfare that is grounded in the conventional economic paradigm makes them of high interest to researchers and policy-makers.

The most commonly used measures of life evaluation refer to "life as a whole" or some similar over-arching construct. However, in addition to global judgements of life as a whole, it is also possible for people to provide evaluations of particular aspects of their lives such as their health or their job. In fact, there is good evidence that a strong relationship exists between overall life evaluations and evaluations of particular aspects of life. One of the most well documented measures of life evaluation – the *Personal Wellbeing Index* – consists of eight questions, covering satisfactions with eight different aspects of life, which are summed using equal weights to calculate an overall index (International Wellbeing Group, 2006). Similarly, Van Praag, Frijters and Ferrer-i-Carbonell (2003) use panel data from the German Socio-Economic Panel to estimate overall life satisfaction as a function of satisfaction with six specific life domains (job satisfaction, financial satisfaction, house satisfaction, health satisfaction, leisure satisfaction and environmental satisfaction), while controlling for the effect of individual personality. These approaches are important because they establish that evaluations of specific aspects of life have a meaningful relationship with overall life evaluations; this therefore suggests that the scope of life evaluations covered in these guidelines needs to encompass specific as well as general measures.

Affect

Affect is the term psychologists use to describe a person's feelings. Measures of affect can be thought of as measures of particular feelings or emotional states, and they are typically measured with reference to a particular point in time. Such measures capture how people experience life rather than how they remember it (Kahneman and Krueger, 2006). While an overall evaluation of life can be captured in a single measure, affect has at least two distinct hedonic dimensions: positive affect and negative affect (Kahneman et al., 1999; Diener et al., 1999). Positive affect captures positive emotions such as the experience of happiness, joy and contentment. Negative affect, on the other hand, comprises the experience of unpleasant emotional states such as sadness, anger, fear and anxiety. While positive affect is thought to be largely uni-dimensional (in that positive emotions are strongly correlated with each other and therefore can be represented on a single axis of measurement), negative affect may be more multi-dimensional. For example, it is possible to imagine at one given moment feeling anger but not fear or sadness.

The multi-dimensional nature of affect raises an interesting question about the relationship of affective states to life evaluation. Life evaluations are uni-dimensional in that different experiences can be rated unambiguously as better or worse. Kahneman et al. (1999), argues for the existence of a "good/bad" axis on which people are able to place experiences based on the emotional states they are experiencing. In effect, he argues, people are able to make an overall judgement about the net impact of their affective state at a particular point in time. In principle, this is the same process that is involved in forming life evaluations from remembered affective states. Kahneman's point is that affective states can be compared, and that one can therefore reasonably aggregate measures of current affect. For this reason, affect measures are sometimes reported in terms of affect balance, which captures the net balance between positive and negative affect (Kahneman and Krueger, 2006).

The measurement of affect poses different challenges to the measurement of life evaluation. It is difficult to ask people to recall affective states in the past, since responses will be affected by recall biases such as the peak/end rule mentioned above. The gold standard for measuring affect is the experience sampling method (ESM), where participants are prompted to record their feelings and perhaps the activity they are undertaking at either random or fixed time points, usually several times a day, throughout the study period, which can last several weeks. To maximise response rates and ensure compliance throughout the day, electronic diaries are often used to record the time of response. While the ESM produces an accurate record of affect, it is also expensive to implement and intrusive for respondents.

A more viable approach is the use of the day reconstruction method (DRM), in which respondents are questioned about events from a time-use diary recorded on the previous day. Research has shown that the DRM produces results comparable with ESM, but with a much lower respondent burden (Kahneman et al., 2004). Experience Sampling, the DRM and similar methods for collecting affect data in time-use studies allow for analysis that associates particular affective states with specific activities. It is also possible to collect affect data in general household surveys.[4] However, affect measures collected in general household surveys lose some detail due to the need to recall affect (even if only what affective states the respondent experienced on the previous day) and also cannot easily capture information linking affect to particular activities.

Eudaimonia

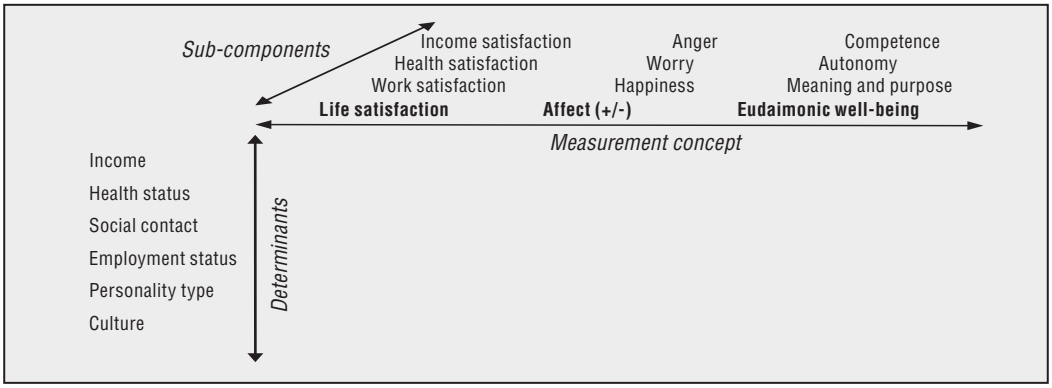
In addition to life evaluations and affect, which focus on a person's experiences (current or recalled), some definitions of subjective well-being found in the psychological literature include other aspects of a person's psychological processes as well. In particular, there is a substantial literature focused on the concept of good psychological functioning, sometimes also referred to as "flourishing" or "eudaimonic" well-being (Huppert et al., 2009; NEF, 2009; Clark and Senik, 2011; Deci and Ryan, 2006). Eudaimonic well-being goes beyond the respondent's reflective evaluation and emotional states to focus on functioning and the realisation of the person's potential. In developing the questionnaire on psychological well-being for the European Social Survey, for example, Huppert et al. (2009) characterise the "functioning" element of well-being as comprising autonomy, competence, interest in learning, goal orientation, sense of purpose, resilience, social engagement, caring and altruism. Eudaimonic conceptions of subjective well-being thus differ significantly from the evaluative and affective components in that they are concerned with capabilities as much as with final outcomes and thus have a more instrumental focus. Because measuring eudaimonia draws on both psychological and humanist literature, which identifies key universal "needs" or "goals", the approach represents a useful response to the criticism that the measurement of subjective well-being is "happiology", or built purely on hedonistic philosophy, and also aligns itself with many people's perceptions of what it is important to value in life.

While there is now a general consensus on the distinction between life evaluations and affect, the conceptual structure of eudaimonic well-being is less well fleshed out. It is not clear, for example, whether eudaimonic well-being describes a uni-dimensional concept in the sense of life evaluation, or whether the term is used to cover a range of different concepts. It is, however, clear that eudaimonic measures of well-being capture important aspects of people's subjective perceptions about their own well-being that are not covered by either life evaluations or affect. For example, having children has a negligible (or even mildly negative) correlation with average levels of life evaluation (Dolan, Peasgood and White, 2008), while child care (even of one's own children) is associated with relatively low levels of positive affect (Kahneman et al., 2004). This conflicts with the intuitive assumption that children, at least for those who choose to have them, contribute in some way to their parent's well-being. Indeed, people with children report much higher average levels of meaning or purpose in their lives than other respondents (Thompson and Marks, 2008).

The relationship between life evaluation, affect and eudaimonia

Life evaluation, positive and negative affect and eudaimonic well-being are all conceptually distinct. However, it is helpful to have a conceptual model of how they might relate to each other. Figure 1.1 provides a simple model of the different elements of a measurement framework for subjective well-being. The model emphasises three dimensions involved in the measurement of subjective well-being. These are: 1) the measurement concept; 2) the sub-components of well-being; and 3) determinants. It should be noted that the list of determinants and sub-components in Figure 1.1 is illustrative rather than exhaustive. The purpose of the conceptual model presented in Figure 1.1 is not to provide a comprehensive framework covering all possible elements of subjective well-being. Rather, it is intended to serve as an organising framework for thinking about the scope of the topics covered by these guidelines. This is necessarily focused on a narrower range of measures than might be found in an academic survey of

Figure 1.1. **A simple model of subjective well-being**



human well-being, and reflects the topics most likely to be of relevance for official statistics and policy-making. There is also a significant bias towards those concepts that underpin the measures traditionally used in large-scale surveys.

Figure 1.1 outlines the various elements of a simple measurement framework for subjective well-being. However, it is also useful to briefly review the empirical relationship between the three types of measures. There is extensive evidence on the relationship between measures of affect and overall measures of life evaluation. Diener, Kahneman, Tov and Arora (in Diener, Helliwell and Kahneman, 2010) show that there is a high correlation (0.82) across countries between the most commonly used average measures of life evaluation, but a much lower correlation (0.55-0.62) between average affect balance and either of two life evaluation measures (life satisfaction and the Cantril Ladder). Similarly, at the individual level, Kahneman and Krueger (2006) report only a moderate correlation (0.38) between life satisfaction (an evaluative measure) and net affect.

There is also a body of evidence on the empirical relationship between eudaimonic well-being and the other aspects of subjective well-being, which suggests that this correlation is smaller than in the case of the relationship between affect and life evaluations. Clarke and Senik (2011), for example, report a correlation between life satisfaction and four different aspects of eudaimonic well-being of between 0.25 and 0.29. Diener et al. (2009) report a correlation of 0.62 ($N = 563$, $p < 0.001$) between their Psychological Well-Being Scale and the evaluative Satisfaction with Life Scale, and correlations of 0.62 and 0.51 respectively between the Psychological Well-Being Scale and the positive and negative subscales of the Scale of Positive and Negative Experience ($N = 563$, $p < 0.001$ in all cases). Huppert and So (2009) found a correlation of 0.32 between flourishing and life satisfaction in European Social Survey data. Among the European Social Survey sample overall, 12.2% met the criteria for flourishing, and 17.7% met the criteria for high life satisfaction, but the percentage for both flourishing and high life satisfaction was 7.2%.

Table 1.1 gives the correlations between individual measures of life evaluation derived from the Gallup World Poll (life satisfaction), positive affect, negative affect and eudaimonic well-being (purpose) across 362 000 respondents in 34 OECD countries. The correlation is highest between the two measures of affect, at -0.3855, and lowest between purpose and negative affect, at -0.091. Life satisfaction has a correlation of about 0.23 with both measures of affect, and of 0.13 with purpose. While all the coefficients in Table 1.1 show the expected sign and all are significant at the 0.1% level, none of the measures have a correlation near 1, indicating that the different measures capture different underlying phenomena.

Table 1.1. **Correlation coefficients among purpose, life satisfaction, positive affect and negative affect at the individual level, 2006-10**

	Purpose	Life satisfaction	Positive affect	Negative affect
Purpose	1.000			
Life satisfaction	0.134	1.000		
Positive affect	0.142	0.229	1.000	
Negative affect	-0.091	-0.231	-0.3855	1.000

Note: The precise measures used are the so-called "Cantril Ladder" for life satisfaction, an "important purpose" in life for purpose, and the sum of "yes" responses to smiled yesterday, experienced joy yesterday, and was well rested yesterday for positive affect and an equivalent index based on experience of sadness, worry and depression for negative affect.
Source: Gallup World Poll.

2. The quality of subjective well-being measures

Quality is crucial to any statistical measure. Unless data captures the concept being measured with a sufficient degree of accuracy to draw reasonable inferences from it, there is little point in collecting it. This is particularly true for official statistics, which are expected to be of the highest quality. As the *United Nations Fundamental Principles of Official Statistics* states, "official statistics provide an indispensible element in the information system of a society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation" (OECD, 2008). It is therefore essential that decisions about the measurement of subjective well-being through official statistics are solidly grounded in a clear understanding of the reliability and validity of such measures.

The *Quality Framework and Guidelines for OECD Statistical Activities* (OECD, 2008) sets out the OECD's approach to dealing with issues of statistical quality. Under the *Framework*, quality is defined as "fitness for use" in terms of user-needs. The ultimate benchmark as to the quality of statistics is essentially whether they meet the needs of the user in terms of providing useful information. Because users must often make decisions about a course of action whether or not statistical information is available, a focus on "fitness for purpose" may involve accepting the use of data that is less than perfectly accurate provided that the data is of sufficient quality that it improves rather than detracts from the quality of decision-making.

Evaluating a concept as broad as "fitness for purpose" is challenging. For this reason, the *Framework* identifies seven dimensions of statistical quality. These seven dimensions define the characteristics of high-quality data and provide a structured way of assessing the quality of a particular set of statistics. The seven dimensions of quality are:

- *Relevance,* i.e. the degree to which data serves to address the purposes for which they are sought by users.
- *Accuracy,* i.e. the degree to which data correctly estimate or describe the quantities or characteristics they are designed to measure.
- *Credibility,* i.e. the confidence that users place in statistics based on their image of the data producer.
- *Timeliness,* i.e. the length of time between the availability of data and the phenomenon or event that the data describe.
- *Accessibility,* i.e. how readily data can be located and retrieved by users.

- *Interpretability,* i.e. the ease with which the user can understand and properly use and analyse the data.
- *Coherence,* i.e. the degree to which the data is mutually consistent with other similar measures and logically integrated into a system of statistics.

These seven criteria, along with the more general principle of cost effectiveness in producing/collecting such data, provide the OECD's overall framework for assessing statistical quality. However, most of these criteria relate to how statistics are measured and collected rather than what is collected. For the purposes of these guidelines, the concern is more narrowly focused on what should be collected rather than the more general principles of how an official statistical agency should operate. Thus, the main focus for assessing the quality of measures of subjective well-being will be the principles of relevance, accuracy and, to a lesser degree, coherence.

3. The relevance of measures of subjective well-being: Why are they important?

It is important to be clear about why subjective well-being should be measured. Official statistics are produced to meet the needs of policy-makers in planning and assessing the impact of policy decisions, and to inform the general public about the state of society. Academics and the media are also important users of official statistics, contributing to a better understanding of society and informing the public and decision-makers. The demand for official statistics is thus, ultimately, a derived demand; statistics are collected because they are of use to someone, rather than for their own sake.

The principles of official statistics generally reflect the view that information is collected only when there is good reason and for a clear purpose. The OECD framework for data quality identifies relevance as the first of the seven key dimensions of quality. Relevance implies that the value of data "is characterised by the degree to which that data serves to address the purposes for which they are sought by users" (OECD, 2003). Similarly, the *United Nations Fundamental Principles of Official Statistics* asserts that the role of official statistical agencies is to compile and make available "official statistics that meet the test of practical utility… to honour citizens' entitlement to public information".

There are sound ethical and practical reasons why official statistical agencies insist on having a clear understanding of the uses of any proposed statistical measures. Many official statistical agencies have the power to compel responses from respondents. That is, respondents are legally required to provide information when approached by a national statistical agency. The corollary of such authority is the requirement for national statistical offices to use data responsibly. From an ethical standpoint, only information that is sufficiently important to justify the intrusion into respondents' lives should be collected. The International Statistical Institute's *Guidelines on Professional Ethics* notes that:

> *Statisticians should be aware of the intrusive potential of some of their work. They have no special entitlement to study all phenomena.*

Over and above this ethical concern is also a practical concern. Even if compliance is legally mandated, the quality of compliance depends heavily on preserving a good relationship between respondents and the official statistical agency. This, in turn, is undermined if the statistical agency cannot articulate why the data being collected is important and how it will be used.

Official statistical agencies are also under increasing resource pressures. This takes the form of both budget cuts, which preclude collecting all the information for which there is a potential demand, and issues of response burden. Even where funding exists to collect information, official statistical agencies must be careful not to over-burden respondents and jeopardise the good will on which high-quality responses depend. Because of this, collecting measures of subjective well-being will have an opportunity cost in terms of other data that will not be collected in order to produce such measures. If subjective well-being measures are to be included in official statistics, therefore, it is essential to be clear about how they will be used.

It is also important to be clear about how subjective well-being measures will be used for purely technical reasons. The field of subjective well-being covers a wide range of different concepts and measures. Choosing which measures should be the focus of collection efforts requires knowing what the measures will be used for. Different measures of subjective well-being will be better suited to different purposes, and it is therefore important that these guidelines identify the right measures needed given the core policy- and public-uses for the data.

The intended use for measures of subjective well-being also affects judgements about the validity of such measures. No statistical measure captures the concept it is intended to measure perfectly. Whether any particular measure can be considered valid, therefore, ultimately involves a judgement about whether the quality of the measure is sufficient to support its intended use. A measure that is valid for one purpose may not be valid for other purposes. For example, a measure could provide valid information about the distribution of outcomes within a country, but be subject to significant bias due to cultural or linguistic factors. While this would be a significant limitation if the intended use of the data is to rank countries compared to each other, it is less important for purely domestic uses.

Measures of subjective well-being have a wide variety of potential uses and audiences. For the purposes of these guidelines it is useful to classify the possible uses of subjective well-being measures under a general framework. The following framework identifies four main ways in which measures of subjective well-being are used. In particular, they can:

- *Complement other outcome measures.*
- Help *better understand the drivers of subjective well-being.*
- Support *policy evaluation and cost-benefit analysis*, particularly where non-market outcomes are involved.
- Help in *identifying potential policy problems.*

Complement other outcome measures

Measures of subjective well-being provide an alternative yardstick of progress that is firmly grounded in people's experiences. These subjective measures may differ in important respects from the picture provided by more conventional metrics that focus on access to resources. This is desirable, since if measures of subjective well-being duplicated the picture provided by other social and economic indicators, there would be few additional gains in using them.[5] In particular, being grounded in peoples' experiences and judgements on multiple aspects of their life, measures of subjective well-being are uniquely placed to provide information on the net impact of changes in social and economic conditions on the perceived well-being of respondents, taking into account the

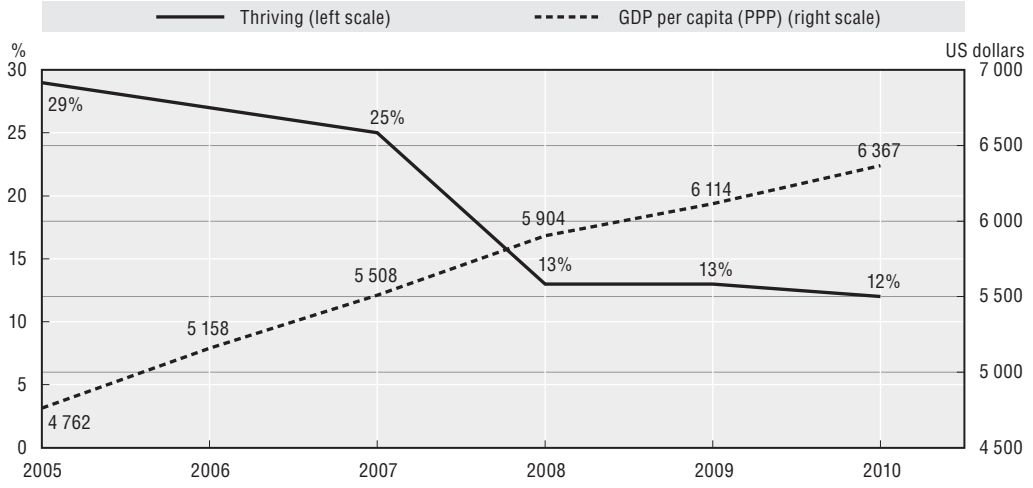
Box 1.1. **Subjective well-being, GDP growth and the "Arab Spring"**

For policy-makers, measures of subjective well-being are valuable as an indicator of progress when they can alert them to issues that other social and economic indicators might fail to identify. One recent example where measures of subjective well-being clearly demonstrate their ability to capture important elements of well-being not captured by more traditional measures is the decline in country-average measures of subjective well-being that occurred in Egypt and Tunisia in the years leading up to 2011, a decline that contrasts with the much more favourable evolution of GDP data. For example, Tunisian real GDP per capita increased from USD 8 891 in 2008 to USD 9 489 in 2010, i.e. a real gain of around 7%. However, the proportion of the population indicating a high level of satisfaction with their life as a whole fell from 24% to 14% over the same period (Gallup, 2011). Egypt (shown in the picture below) shows a similar pattern from 2005 to 2010, with a real gain in GDP per capita of around 34% and a decline in the share of respondents classified as "thriving" by almost half.* This illustrates how subjective perceptions can provide information on very significant outcomes in societies that other conventional indicators such as GDP growth do not provide.

* "Thriving" is a composite measure of subjective well-being calculated by the Gallup Organisation. It is based on answers to the Cantril ladder measure of life satisfaction for life at the moment and how people expect life to be in five years.

Figure 1.2. **Trends in subjective well-being and GDP in Egypt: 2005-10**

Recent trends in percentage "thriving" and GDP per capita (PPP)



Source: Subjective well-being data are from Gallup. GDP per capita (PPP) estimates are from the *International Monetary Fund's World Economic Outlook Database*.

impact of differences in tastes and preferences among individuals. An example of how these measures can change perceptions about progress in individual countries is provided by Box 1.1, in respect of the "Arab Spring".

In addition to providing information on aggregate changes at the national level, measures of subjective well-being can also provide a picture of which groups in society are most (dis)satisfied or experience the best or worse life. Again, because measures of subjective well-being capture the impact of taste and aspirations as well as the distribution of other life circumstances, such measures provide useful additional information for policy-makers in situations where comparisons are made across sub-groups of the

population. Migrants, for example, may be more motivated than the rest of the population by income relative to other factors (Bartram, 2010), as this is a primary motive for their decision to move abroad. An attempt to assess migrant well-being compared to the rest of the population is therefore challenging, given that there is good reason to believe that there will be systematic differences in the importance that the two groups attach to different aspects of quality of life. Because measures of subjective well-being incorporate the impact of the different weights that various people may attach to the different aspects of their quality of life, they have the potential to add an important dimension to analysis in situations involving comparisons between population groups.

The final policy use of measures of subjective well-being in the context of measuring progress is for cross-country comparisons of aggregate measures of subjective well-being, such as those included in *How's Life?* (OECD, 2011). Due to the impossibility of performing controlled experiments across countries, cross-country comparisons of subjective well-being outcomes are one way to learn about the strengths and weaknesses of different policies. Because measures of subjective well-being are sensitive to a different range of drivers than are other social and economic indicators, they can provide additional information about the consequences of a particular policy. A crucial issue in using measures of subjective well-being in this way, however, is the degree to which cross-cultural comparisons of measures of subjective well-being are valid. This issue is considered in more depth later in this chapter.

The interest of the general public and the media in using measures of subjective well-being as complements of measures of progress is generally similar to that of policy-makers. For these users, the key contributions that subjective well-being measures can potentially make are in highlighting how different groups fare compared to each other, what can be learned from the experiences of other countries, and perhaps whether things are getting better or worse overall – all of which are of potential interest to the general public and the media.

Better understand the drivers of subjective well-being

The second major use of subjective well-being measures is to contribute to a better understanding of the drivers of well-being at an individual level. If it is accepted that measures of subjective well-being are valid, and that they accurately capture the concepts that they claim to measure – an overall evaluation of life or the experienced moods and emotions of an individual over a period of time – then it follows that such measures can be used to provide information on the relative contribution of different factors and circumstances to a person's well-being – albeit with some noise due to both measurement error and the fact that a person's subjective perception of their well-being is not necessarily quite the same thing as their overall well-being (for examples see Dolan, Peasgood and White, 2008; Helliwell and Wang, 2011; Boarini, Comola, Smith, Manchin and De Keulenaer, 2012).

Measures of subjective well-being can be used to help identify what factors are critical aspects of people's well-being. In particular, such measures can be used to test intuitions about what factors matter most to people. This is potentially important to the broad agenda of measuring progress, since it provides an empirical way to test whether the outcomes used to measure progress align well with the factors that determine people's perceptions of their well-being. Although people's subjective perceptions are not necessarily equivalent to overall well-being for a number of reasons,[6] measures of subjective well-being are unique in that they provide a relatively robust empirical source of

information on such preferences, especially when non-market outcomes are involved. Without using subjective views of what matters the most to people, we would be left to essentially *a priori* judgements and anecdotal focus group research.

Subjective well-being measures are, however, unique in that they provide a relatively robust empirical source of information on what affects how people feel about their lives, which is an important component of overall well-being. By examining the level of subjective well-being *actually achieved* as a result of different decisions or approaches, policy-makers and individuals can better understand what matters to people on an empirical (rather than anecdotal) level. For example, subjective measures can be used to test more specific hypotheses about what aspects of policy are most important to people. Halpern (2010), for example, refers to an instance where the Merseyside police, in the United Kingdom, used data on how satisfied members of the public were with the service provided by the local policy, alongside more traditional performance measures on crimes committed and offence resolutions. In contrast to the expected hypothesis – which was that minimising the response time from the police was of crucial importance for public satisfaction – the evidence showed that it was much more important that police arrived when they said they would. For minor issues not involving safety, what mattered was punctuality rather than speed.

Going beyond just identifying what matters for well-being, measures of subjective well-being can assist in developing a better understanding of the trade-offs between different outcomes. Many policy problems require taking a decision about how to compare two fundamentally different types of outcome (see Box 1.2). Dolan and White (2010) note that these types of issue characterise many attempts to encourage "joined-up government", where there is a need for different government agencies to consider the costs and benefits of a particular intervention not just on their primary outcome of concern, but also in terms of how these affect the outcomes of other government agencies.

Because measures of subjective well-being can capture the combined effect of all different changes in life circumstances on an individual's perception of their well-being in a single measure, they can be used as a common metric for assessing the relative impact of fundamentally different outcomes. For example, Ferrer-i-Carbonell and Frijters (2004) use measures of overall satisfaction with life and satisfactions with specific outcome domains to assess the relative weights to attach to different outcome areas. Comparing the magnitude of the impact of health satisfaction on overall life satisfaction with the impact associated with housing satisfaction gives a way of quantifying the relative importance of each dimension within a particular sample, given where they started on each measure.[7] Similarly, Di Tella, Oswald and Maculloch (2003) used the coefficients from a regression on life satisfaction to investigate the inflation/unemployment trade-off. While the so-called "misery index" weights the unemployment rate and inflation rate equally as indicators of the negative impact of macroeconomic outcomes, Oswald and Maculloch's analysis suggests that the impact of unemployment on subjective well-being is significantly greater than that of inflation.

Policy evaluation and cost-benefit analysis

The third main use of subjective well-being measures is to assist in the evaluation of policies. This includes both the direct use of measures of subjective well-being in formal policy evaluations as well as the more indirect – but possibly more important – role that they can play in cost-benefit analysis.

Box 1.2. **Using measures of subjective well-being to value life events**

People intuitively compare different life events on a daily basis and make judgements about how bad or good things might be. However, trying to put a number on the relative magnitude of the impact of different life events such as marriage or divorce, on a person's well-being – much less a monetary value – might seem counter-intuitive to many people. Nonetheless, such values are of potentially high interest from the perspective of thinking about how much to invest in preventing or encouraging a particular outcome.

Measures of subjective well-being provide a relatively straight-forward way of comparing the relative impact of fundamentally different life events in a quantitative way and, based on this, assigning such events a monetary value. Clark and Oswald (2002) present a method for valuing life events and, although the literature on using measures of subjective well-being to value life events has expanded significantly since 2002, the basic methodology remains largely unchanged. Consider the results below from a regression of a range of possible determinants of subjective well-being against life satisfaction (Boarini et al., 2012). The coefficients for the (base two) logarithm of household income, being married, and being unemployed are shown, and express the change in life satisfaction (on a scale of 0 to 10) associated with a doubling of income, being married, or being unemployed, respectively, holding all else constant.

Event	Coefficient
Log household income	0.1482
Married	0.2584
Unemployed	-0.4643

Using these coefficients, it is possible to calculate the relative impact of being married compared to being unemployed on life satisfaction as 0.2584/0.4643 = 0.5565. Or, put more simply, being unemployed has almost twice the impact on life satisfaction as does being married.

Going beyond this, the monetary value of being married or being unemployed can be calculated by comparing the relevant coefficients with that associated with the coefficient for household income. Using the values presented above, the coefficient on being married is 0.2584/0.1482 = 1.7435 times larger than the impact of a doubling of household income. For a person with a household income equal to the OECD per capita household disposable income (USD 17 286 at PPP, 2008), this is equivalent to 1.7435 × USD 17 286 = USD 30 138. For unemployment the comparable value is 2.930 × USD 17 286 = USD 50 647.

These values are intended to illustrate the techniques involved, and need to be treated with caution. In particular, better measures would use panel data to capture the causal relationship (as do Clark and Oswald) rather than just correlation, and need to consider any potential biases in the data as well as the structure of the regression equations used to calculate the coefficients (Fujiwara and Campbell, 2011).

In formal policy evaluations, measures of subjective well-being can complement other social and economic indicators as a measure of the outcomes achieved by a policy. Here, as is the case with monitoring the progress of entire communities, measures of subjective well-being can add additional information over and above that captured by more traditional indicators. For some initiatives – where the impact on the subjective

experiences of the population is the main object of the programme – measures of subjective well-being may even be suitable as the primary metric for assessing the programme's success.

Many policy evaluations already include subjective measures of client satisfaction and questions on the respondent's perceptions of what elements of the programme were most valuable. More general measures of overall subjective well-being, however, have some significant advantages over and above these more focused measures. Most importantly, measures of subjective well-being provide information on the actual impact of an initiative on the respondent's subjective well-being, rather than the impact that the respondent consciously identifies. These values can differ because people's judgements about the impact of a programme may be influenced by the fact that they have participated in the programme (i.e. they might be more prone to assign the cause of any recent changes in their well-being to the programme rather than to other factors, knowing that this is what he/she is being asked about). Also, people may not be aware of all of the various feedback loops via which a policy programme affects them. For example, in evaluating an active employment programme, respondents might consider the direct effect on their well-being of both having a job and gaining additional income, but not the flow on well-being that would stem from changes in their time-use due to longer commuting. Because measures of subjective well-being can capture the overall impact of a change on life circumstances, without requiring a cognitive judgement by the respondent on which causal pathways are being asked about, such measures provide useful additional information on the overall impact of a programme.

In some cases, measures of subjective well-being can be better than conventional cost-benefit analysis at treating non-monetary outcomes. Examining the relative costs and benefits of a proposal is relatively straight-forward when the proposal is aimed at strictly economic outcomes, and the costs and benefits of the proposal can be obtained from the relevant market prices. However, where the aim of a proposal is to achieve outcomes that do not have an obvious market price, it is much more challenging to obtain meaningful values for analysing the relevant costs and benefits. Because much government policy is concerned with market failures, many government policies are correspondingly focused on achieving non-market outcomes.

The traditional economic approaches to cost-benefit analysis for non-market outcomes depend on either revealed preference or contingent valuation techniques to estimate "prices" for such outcomes. A revealed preference approach involves calculating values based on the shadow prices implied by observed behaviour, while contingent valuation techniques calculate values based on the "willingness to pay" for the outcome in question, as expressed by respondents to a hypothetical question in a survey. Clarke and Oswald (2002) note that measures of subjective well-being can provide the framework for such valuations by comparing the impact of a particular outcome on subjective well-being with the impact of a change in income on subjective well-being. By making such a comparison, it is possible to calculate the amount of money required to achieve the same increase or decrease in well-being as that caused by the outcome under assessment.

There is good reason to believe that, in several circumstances, measures of subjective well-being have advantages over both revealed preference and contingent valuation for the purposes of cost-benefit analysis (see Box 1.3). An obvious advantage is that many measures of subjective well-being – such as overall life satisfaction – are relatively easy and

Box 1.3. *The Green Book* and life satisfaction

The Green Book is the formal guidance from the Treasury of the United Kingdom to other UK government agencies on how to appraise and evaluate policy proposals. The current edition of *The Green Book* dates to 2003, and provides advice on how officials should provide justification for a proposed government intervention, set objectives for the proposal, appraise the various options, and evaluate the effectiveness of the final action that results. In July 2011, *The Green Book* was updated to reflect the results of a review of valuation techniques for social cost-benefit analysis jointly commissioned by the Treasury and the Department for Work and Pensions (Fujiwara and Campbell, 2011). The review specifically focuses on the contribution that can be played by measures of subjective well-being – particularly life satisfaction – alongside more traditional approaches to cost-benefit analysis. In summarising the conclusions of the review, *The Green Book* states (p. 58):

> *A newer, "subjective well-being approach" has been gaining currency in recent years. The "life satisfaction approach" looks at people's reported life satisfaction in surveys such as the ONS's Integrated Household Survey, which began including questions on respondents' subjective well-being in April 2011. The life satisfaction approach uses econometrics to estimate the life satisfaction provided by certain non-market goods, and coverts this into a monetary figure by combining it with an estimate of the effect of income on life satisfaction.*
>
> *At the moment, subjective well-being measurement remains an evolving methodology and existing valuations are not sufficiently accepted as robust enough for direct use in Social Cost-benefit Analysis. The technique is under development, however, and may soon be developed to the point where it can provide a reliable and accepted complement to the market based approaches outlined above. In the meantime, the technique will be important in ensuring that the full range of impacts of proposed policies are considered, and may provide added information about the relative value of non-market goods compared with each other, if not yet with market goods.*

While the amendment to *The Green Book* stops short of fully endorsing the use of life satisfaction measures for use in formally evaluating government programmes, the decision to make an interim amendment in itself signals strongly the importance that UK central agencies attach to obtaining improved measures of the value of non-market outcomes.

cheap to collect. However, there are also more substantive methodological advantages that may be associated with using measures of subjective well-being in this way. Revealed preference relies on strong assumptions about people's ability to know how an outcome will affect them in the future, and on the assumptions that markets are in equilibrium. Diener, Lucas, Schimmack and Helliwell (2009) note that for market prices for houses to reflect the disutility of airport noise accurately would require that house purchasers are able to forecast how much the noise will impact them before buying the house. Similarly, in this example, it is difficult to disentangle the differences in house prices due to noise from differences in other aspects of house quality.

Contingent valuation also relies strongly on people's ability to make accurate judgements about how something will make them feel in the future. Dolan and Peasgood (2006) note that people have difficulty imagining how good or bad different circumstances are actually going to be. Indeed, the "willingness to pay" surveys commonly used for contingent valuation are, to a large degree, measures of the subjective well-being associated with a hypothetical scenario. Using measures of subjective well-being to calculate the costs based on the actual impact of different life circumstances on subjective

well-being removes the hypothetical element from the equation. In addition, contingent valuation surveys tend to produce very different estimates of the value of outcomes for people at different points on the income distribution. This tends to result in either weighing the desires of the rich more heavily than the poor when assessing the costs and benefits associated with the proposal under consideration or taking account of the marginal utility of income in calculating the final cost. The latter approach is difficult in the absence of robust estimates of the marginal utility of income (Dolan and White, 2007).

Identifying potential policy problems

An important feature of measures of subjective well-being is their ability to provide an insight into human behaviour and decision-making. In particular, measures of subjective well-being can help researchers to understand the difference between the *ex ante* beliefs that people hold about their future well-being (which form the basis for decisions) and the *ex post* outcomes that people achieve in terms of their subjective well-being. A better understanding of these issues is important both for policy-makers and for the broader public. Policy-makers have an interest in understanding why people make the decisions that they do, because much public policy involves dealing with the consequences of systematic poor decision-making by individuals. Similarly, businesses and the general public have an interest in understanding how people's subjective well-being shapes their behaviour.

One way in which measures of subjective well-being are useful to businesses and the broader public is by providing information on the characteristics of good places to live and work. There is clear evidence that subjective well-being predicts future behaviour. Clark (2001), for example, has shown that measured job satisfaction predicts the probability of an employee going on to leave their job. Thus businesses might well have an interest in the measured job satisfaction of their employees and in understanding the determinants of job satisfaction.

Measures of subjective well-being can also help shed light on various biases in the way people make decisions. Although people are generally able to predict whether events are likely to be pleasant or unpleasant, Wilson, Gilbert and colleagues have described various ways in which affective forecasting can be biased or faulty, particularly with regard to the intensity and duration of emotional reactions to future events (e.g. Wilson, Wheatley, Meyers, Gilbert and Axsom, 2000; Wilson and Gilbert, 2006). Kahneman et al. (2006), for example, show that people are prone to over-estimate the impact of income gains on their life satisfaction. When evaluating those factors that people expect to contribute to a positive mood, people tend to focus on conventional achievements, thus over-estimating the role of income relative to other factors. By way of contrast, other activities that are less commonly used as a reference for conventional measures of status get under-estimated with respect to their impact on subjective well-being. Commuting, for example, has been found to have a strong negative impact on both measures of affect (Kahneman et al., 2006) and life evaluations (Frey and Stutzer, 2008). This suggests that people may be prone to over-estimating the positive impact of, for example, a new job with a higher salary but a longer commute.

Faulty affective forecasting is significant in this context because it suggests that decisions reflected in market choices will not always serve to maximise subjective well-being in practice. Individuals may have a substantial interest in better understanding the factors affecting the level of well-being that they actually achieve. Hence, a sound evidence base derived from measures of subjective well-being is of potential interest to the general public. There are also direct policy applications for better understanding the

human decision-making process and the various biases and heuristics involved in it. Consider the case of policy options that incorporate a "default" option, for example, workplace retirement schemes that are set up on a basis of either "opt in" clauses, where a new employee does not join the scheme unless he/she ticks a box to join, or "opt out" clauses, where the reverse is the case. The fact that people respond differently depending on which default is selected – despite the fact that in neither case is there any compulsion – has raised policy interest in the idea of "libertarian paternalism", which focuses on achieving better outcomes be setting policy defaults to influence people's behaviour in positive directions. Dolan and White (2007) note that information on subjective well-being can be used to help set policy default options appropriately, by indicating which default options contribute most to subjective well-being.

4. The accuracy of subjective well-being measures

Accuracy is concerned with whether the measure in question accurately describes the qualities of the concept being measured. This, in turn, is usually assessed in terms of reliability and validity. Reliability concerns the extent to which a measure yields consistent results (i.e. whether it has a high signal-to-noise ratio). Validity, on the other hand, is about the extent to which it actually captures the underlying concept that it purports to measure (i.e. whether it is measuring the right thing). Some degree of reliability is a necessary but not sufficient condition for validity.

Reliability

Reliability is a fundamental component of accuracy. For any statistical measure it is desirable that the measure produce the same results when carried out under the same circumstances. This is essential if the measure is to be able to be used to distinguish between changes in the measure due to a genuine change in the condition being measured as opposed to changes that simply represent measurement error. While no statistical measure is completely reliable, it would be of concern if measures of subjective well-being performed significantly worse than other commonly-used measures.

There are two main ways to measure reliability. Internal consistency reliability concerns the extent to which different items on an overall scale or measure agree with one another, and is assessed through examination of inter-item correlations. If the correlation between the two items is high, this suggests that the two measures capture the same underlying concept. On the other hand, if the correlation is low, it is not necessarily the case that both measures are poor but at least one of them must be.

The second approach involves looking at test-retest reliability, where the same question is administered to the same respondent more than once, separated by a fixed period of time. Test-retest reliability places a lower bound on the overall reliability of the measure, but not an upper bound. For example, a low test-retest score could indicate that a measure lacks reliability, but it could also be associated with a high level of actual reliability and a genuine change in the subject of interest.

Both aspects of reliability have been extensively tested for measures of life evaluation and affect over the past twenty years. There is strong evidence for convergence between different life satisfaction measures. In a meta-review of the reliability and validity of subjective well-being measures, Diener (2011) reports a Cronbach's alpha[8] for multiple item measures of life satisfaction (including the Satisfaction With Life Scale) of between 0.8

and 0.96. A Cronbach's alpha of 0.7 is typically taken to be the threshold of acceptable convergence, and the scores Diener reports indicate a very high degree of convergence between the different questions used in the life satisfaction scales.

Whilst it is not possible to compute Cronbach's alpha for single item measures, other estimation procedures can be employed. Comparisons across countries using different measures of the same construct generally show slightly lower correlations, but are still relatively high given that the scores not only represent different questions, but are also sourced from different surveys. Bjørnskov (2010), for example, finds a correlation of 0.75 between the average Cantril Ladder measure of life evaluation from the Gallup World Poll and life satisfaction as measured in the World Values Survey for a sample of over 90 countries.

Test-retest results for single item life evaluation measures tend to yield correlations of between 0.5 and 0.7 for time periods of 1 day to 2 weeks (Krueger and Schkade, 2008). Michalos and Kahlke (2010) report that a single-item measure of life satisfaction had a correlation of 0.67 for a one-year period and of 0.65 for a two-year period. In a larger study, Lucas and Donellan (2011) estimated the reliability of life satisfaction measures in four large representative samples with a combined sample size of over 68 000, taking into account specific errors. They found test-retest correlations in the range of 0.68 to 0.74, with a mean of 0.72 over a period of one year between reports.

Multiple item measures of subjective well-being also generally do better than single questions on test-retest reliability. Krueger and Schkade (2008) report test-retest scores in the range of 0.83 to 0.84 for a period of 2 weeks to 1 month between tests, with correlations declining to 0.64 at 2 months and to 0.54 over 4 years. The pattern of decline here is as expected, with longer periods of time showing lower reliability due to a higher likelihood that there has been a genuine change in the respondent's circumstances. In another study, the "satisfaction with life" scale (a multi-item measure of life satisfaction) showed a correlation coefficient of 0.56, dropping to 0.24 after 16 years (Fujita and Diener, 2005).

Generally speaking, country averages of measures of subjective well-being show higher levels of stability than do those for individuals. Diener (2011), for example, reports a correlation coefficient of 0.93 for the Cantril Ladder in the Gallup World Poll over 1 year, and a correlation of 0.91 across 4-year intervals.

There is less information available on the reliability of measures of affect and eudaimonic well-being than is the case for measures of life evaluation. However, the available information is largely consistent with the picture for life satisfaction. In terms of internal consistency reliability, Diener et al. (2009) report that their Psychological Well-Being scale has a Cronbach's alpha of 0.86 (N = 568), whilst the positive, negative and affective balance subscales of their Scale of Positive and Negative Experience (SPANE) have alphas of 0.84, 0.80 and 0.88, respectively (N = 572, 567 and 566).

In the case of test-retest reliability, which one might expect to be low in the case of momentary affect, but higher in the case of longer-term affective experiences, Krueger and Schkade (2008) report test-retest scores of between 0.5 and 0.7 for a range of different measures of affect over a 2-week period. Watson, Clark and Tellegen (1988) report slightly lower scores of between 0.39 and 0.71 for a range of different measures over an 8-week period. The lower scores are recorded by measures of momentary affect, while the upper scores are for questions focusing on affective states over a longer period of time, so the range of scores is consistent with expectations. Diener et al. (2009) meanwhile report a correlation of 0.71 (N = 261) between measures of Psychological Well-Being issued one

month apart, whilst the positive, negative and affect balance measures of the SPANE had coefficients of 0.62, 0.62 and 0.68, respectively (*N* = 261). Clark and Senik (2011) meanwhile report a Cronbach's alpha of 0.63 in their eudaimonia measure, which is derived from items in the well-being module of the European Social Survey (*N* = over 30 000 respondents).

Summary: Reliability

Taken as a whole, test-retest scores for measures of subjective well-being are generally lower than is the case for commonly collected demographic and labour market statistics such as education and income. These variables generally show a test-retest score in the region of 0.9 (Krueger and Schkade, 2008) – although one would of course expect education and income to vary less over very short time periods. However, the scores for measures of subjective well-being are higher than those found for some more cognitively challenging economic concepts and for those that one would expect to be more variable over time. For example, an analysis of expenditure information found test-retest values of 0.6 over a period of one week (Carin, G., D. Evans, F. Ravndal and K. Xu, 2009).

In general, a cut-off at 0.7 is considered an acceptable level of internal consistency reliability for tests based on comparing results when using different measures (Nunnally and Bernstein, 1994; Kline, 2000). By this criterion, the more reliable multi-item measures of subjective well-being, such as the satisfaction with life scale, exhibit high reliability, although they are not as reliable as demographic statistics or educational status. The case for single item measures is more ambiguous, although the analysis of Lucas and Donellan, which has the best measures and largest dataset of any of the studies considered here, suggests that single item measures of life satisfaction also have an acceptable degree of reliability. Looking at country averages, the reliability of life satisfaction measures is generally well above the required threshold for acceptable reliability.

Measures of affect would be expected to have lower levels of reliability than is the case for evaluative measures, simply because moods change more frequently. The available evidence is generally consistent with this, and suggests that affect measures are reliable enough for use. There is less evidence on measures of eudaimonia. Although the Diener/Wirtz Psychological Well-being Scale performs relatively well, this cannot necessarily be generalised to other measures of eudaimonic well-being, and further research is warranted in this area.

Validity

While the reliability of a measure is largely a function of the degree to which it produces consistent results, and is therefore relatively easy to test, validity is more challenging to establish. This is particularly the case for subjective measures. Although there is a broad consensus around the range of concepts to be measured, this does not, in and of itself, mean that establishing the validity of a measure is straight-forward. If a measure is subjective, how can we know whether it is a good measure of the underlying concept?

At this point some precision is required about what is meant by a subjective measure. There are two senses in which one can talk about a subjective measure of something. A subjective measure can refer either to the measure itself, or to the concept being measured. In the first sense, it is possible to have a subjective measure of an objective concept (as in the case of the question, "who do you think is older, John or Marama?"). The measure is subjective in that it seeks a person's opinion, but the subject being measured (the ages of John and Marama) can be objectively verified, i.e. by checking the dates of birth for both John and Marama, different people will provide the same outcome.

When the concept itself is subjective, however, things become more complicated. In the case of the question "how much do you like the colour blue", the concept being measured is itself subjective. There is no way for a person other than the respondent to provide the correct answer. This makes testing the validity of such measures much more challenging than in the first instance.

Measures of subjective well-being are subjective in this second sense, and this means that we cannot compare measures of subjective well-being with objective measures of the same concept in order to reassure us of their validity. However, this does not mean that we cannot meaningfully analyse the validity of measures of subjective well-being at all. There is an extensive psychological literature on the validity of subjective measures, and this literature suggests three types of validity that a good subjective measure should demonstrate:

- Face validity, i.e. do respondents and/or data users judge that the items are appropriate, given what they are told about the assessment objectives?
- Convergent validity, i.e. does the measure correlate well with other proxy measures for the same underlying concept?
- Construct validity, i.e. does the measure perform in the way theory would suggest with respect to the construct being measured?

Face validity

The face validity of measures is important because it can affect both respondent motivation and the acceptability of resulting data to potential users of that data. The face validity of subjective well-being is relatively straight-forward to establish. The standard questions used have a clear intuitive relationship to the concept being measured. It is not a great stretch, for example, to suggest that asking a person whether they experienced sadness during the previous day is a plausible way to find out whether they felt sad during that day. Although it is relatively unusual to ask respondents about face validity directly, there are a number of pieces of evidence that suggest that respondents find questions on subjective well-being easy to understand. For example, the time it takes respondents to reply to questions on subjective well-being is low, with median response times well under thirty seconds for single item questions (ONS, 2011). This indicates that respondents do not generally find such measures conceptually difficult to understand. Cognitive testing by the ONS also supports the view that respondents do not generally find subjective questions difficult or upsetting to answer, nor does the inclusion of such questions negatively impact the response rates to subsequent questions or to the survey as a whole (ONS, 2011; ONS, 2012). Measures of subjective well-being also have low item-specific non-response rates (Rässler and Riphahn, 2006), suggesting that respondents do not find these types of question difficult to answer.

In a large analysis by Smith (2013) covering three datasets (the Gallup World Poll, the European Values Survey and the World Values Survey) and over 400 000 observations, item-specific non-response rates for measures of life evaluation and affect were found to be similar to those for measures of educational attainment, marital status and labour force status. The acceptability of the subjective well-being measures, however, appeared to be higher than that of income, which is commonly collected as part of official statistics, and had an item-specific non-response rate of between 10 and 100 times higher than subjective well-being measures, depending on the country. The results also held when item-specific

non-response rates were broken down by cause (into "don't know" and "refused to answer") and regardless of whether the measure tested was evaluative (life satisfaction or the Cantril Ladder) or affective.

Convergent validity

Convergent validity involves examining whether a measure correlates well with other proxy measures for the same concept. Although measures of subjective well-being are focused on an inherently subjective concept, a range of information can be used as proxy measures for people's subjective states. One option is to look at how ratings from the respondent compare to ratings from other people, such as friends, families, or even the interviewer. Similarly, one can observe the behaviour of the respondent to see whether it is consistent with their reported subjective state. Finally, one can use biophysical measures related to emotion states. All of these approaches have been applied to measures of subjective well-being and provide strong support for convergent validity.

Ratings of a person's subjective well-being from friends and family have been shown to correlate well with self-reports of life satisfaction (Frey and Stutzer, 2002). A review by Pavot and Diener (1993) found correlations of between 0.43 and 0.66 between interviewer ratings and self-ratings, and correlations of between 0.28 and 0.58 between self-reports and other informants, such as friends and families. In a meta-analysis of self-informant ratings, Schneider and Schimmack (2009) found a mean correlation of 0.42 between self-reports of life satisfaction and informant reports. Similarly, for momentary affect, strangers shown a video or pictures of the respondent are able to accurately identify the subject's dominant emotion at a particular point in time (Diener, Suh, Lucas and Smith, 1999). This latter finding also held when the informant was a person whose culture differed fundamentally from that of the respondent.

Subjective assessments of well-being are also reflected in behaviour. People who rate themselves as happy tend to smile more. This applies particularly to so-called "Duchenne" or "unfakeable" smiles, where the skin around the corners of the eye crinkles through a largely involuntary reflex (Frey and Stutzer, 2002; Diener, 2011). There is also good evidence that people act in ways that are consistent with what they say about their subjective well-being, i.e. people avoid behaviours that they associate with a low level of subjective well-being (Frijters, 2000). Diener (2011), summarising the research in this area, notes that life satisfaction predicts suicidal ideation ($r = 0.44$), and that low life satisfaction scores predicted suicide 20 years later in a large epidemiological sample from Finland (after controlling for other risk factors such as age, gender and substance use). Self-reports of job satisfaction have been shown to be a strong predictor of people quitting a job, even after controlling for wages, hours worked and other individual and job-specific factors (Clark, Georgellis and Sanfrey, 1998).

There have been a number of studies that look at the correlation between various bio-physical markers and subjective well-being. Measures of subjective well-being have been shown to be correlated with left/right brain activity (Urry et al., 2004; Steptoe, Wardle and Marmot, 2005). Activity in the left prefrontal cortex of the brain has been shown to be strongly correlated with processing approach and pleasure, while activity in the corresponding part of the right hand side of the brain is correlated with processing avoidance and aversive stimuli. Steptoe, Wardle and Marmot (2005) also investigated the association between the level of the stress hormone cortisol in the bloodstream and self-reported happiness, finding a 32% difference in cortisol levels between people in the

highest and lowest quintiles of happiness. People reporting high levels of subjective well-being also recover more quickly from colds and minor injuries (Kahneman and Krueger, 2006).

There is no clear cut-off point for what constitutes an acceptable level of convergent validity. This is because the measured relationship depends as much on the quality of the proxy variable to which the measure is being compared as it does on the validity of the measure itself. However, the available evidence on convergent validity does allow for claims to be made about subjective well-being measures from the perspective of falsifiability. In particular, if it were found that a plausible proxy measure of subjective well-being showed no or a negative relationship with a measure of subjective well-being, this would be taken as good evidence that the measure in question lacked validity until the finding was over-turned, either because a good explanation for the relationship was found or because the result failed to be replicated. The consistent positive relationship found between measures of life satisfaction and the wide range of proxy measures considered above suggests strongly that such measures can be considered as displaying adequate convergent validity. The evidence for the convergent validity of affective measures is also persuasive. However, there is little to draw on with respect to convergent validity and eudaimonic measures.

Construct validity

Where convergent validity involves assessing the degree to which a measure correlates with other proxy measures of the same concept, construct validity focuses on whether the measure performs in the way that theory would predict. Construct validity concerns itself with whether the measure shows the expected relationship with the factors thought to determine the underlying concept being measured, and with outcomes thought to be influenced by the measure in question. There is an extensive literature relevant to the assessment of the construct validity of measures of subjective well-being. Economists in particular, driven in part by the desire to understand how well such measures perform as a potential measure of utility, have looked in depth at the economic and social drivers of subjective well-being. Meanwhile, psychologists have often explored individual, psychological and psycho-social determinants.

Measures of subjective well-being broadly show the expected relationship with other individual, social and economic determinants. Among individuals, higher incomes are associated with higher levels of life satisfaction and affect, and wealthier countries have higher average levels of both types of subjective well-being than poorer countries (Sacks, Stevenson and Wolfers, 2010). At the individual level, health status, social contact, education and being in a stable relationship with a partner are all associated with higher levels of life satisfaction (Dolan, Peasgood and White, 2008), while unemployment has a large negative impact on life satisfaction (Winkelmann and Winkelmann, 1998. Kahneman and Krueger (2006) report that intimate relations, socialising, relaxing, eating and praying are associated with high levels of net positive affect; conversely, commuting, working, childcare and housework are associated with low levels of net positive affect. Boarini et al. (2012) find that affect measures have the same broad sets of drivers as measures of life satisfaction, although the relative importance of some factors changes.

Further, it is clear that changes in subjective well-being – particularly life evaluations – that result from life events are neither trivial in magnitude, nor transient. Studies have shown that change in income, becoming unemployed, and becoming disabled have a

long-lasting impact on life satisfaction (e.g. Lucas, 2007; Lucas, Clark, Georgellis and Diener, 2003; Diener, Lucas and Napa Scollon, 2006), although there can also be substantial individual differences in the extent to which people show resilience, or are able to adapt to, adversity over time. In the case of negative life experiences, Cummins et al. (2002) note that extreme adversity is expected to result in "homeostatic defeat" – thus, life experiences such as the chronic pain of arthritis or the stress of caring for a severely disabled family member at home can lead to stably low levels of subjective well-being. Similarly, Diener, Lucas and Napa Scollon (2006) describe evidence of partial recovery from the impacts of widowhood, divorce and unemployment in the five years following these events, but subjective well-being still fails to return to the levels observed in the five years prior to these events. Thus although there is evidence of partial adaptation to changes in life circumstances, adaptation is not complete, and the impact of these life events on life evaluations is long-lasting.

Summary: Validity

Over the last two decades an extensive body of evidence has accumulated on the validity of measures of life evaluation and affect. This evidence strongly supports the view that measures of both life evaluation and affect capture valid information. This does not mean that these measures are universally valid or devoid of limitations. However, these limitations do not suggest that measures of subjective well-being should be regarded as not fit for purpose if used with appropriate caveats. The evidence base for eudaimonic measures is less clear. While psychologists have studied concepts related to eudaimonic well-being such as good psychological functioning for some time, it has proved more difficult to pull together a summary of the literature addressing validity in the terms set out above. This does not mean that eudaimonic measures are not valid, but suggests that further work is needed before a definitive position can be taken on the validity of these measures. Table 1.2 provides a summary of the evidence for the validity of subjective well-being outlined above.

Limits of validity

Although the evidence for the reliability, validity and usefulness of subjective well-being measures is strong, like all measures they are not perfect, and there are limitations that need to be considered by both producers and users of subjective well-being data.

Subjective well-being measures have been found to have a relatively high noise-to-signal ratio. For example, in reviewing the evidence, Diener (2011) states that around 60-80% of the variability in life satisfaction scales is associated with long-term factors and that the remaining 20-40% is due to occasion-specific factors and errors of measurement. These occasion-specific factors can include one-off occurrences that affect large numbers of people simultaneously, such as major news events or Valentine's Day (Deaton, 2011), or circumstantial events that may affect individuals' momentary mood prior to the survey (Schwarz and Strack, 2003). Whilst the latter effect should be sufficiently random to wash out of large representative data sets, the former implies that a reasonable number of *days*, as well as people, need to be sampled to reduce the risk of systematic error. This is further supported by work demonstrating that the day of the week (e.g. Taylor, 2006; Helliwell and Wang, 2011), the season (Harmatz et al., 2000) and the weather (e.g. Barrington-Leigh, 2008) can also influence certain subjective well-being measures,[9] although results do tend to be more mixed in these areas.

Table 1.2. **Evidence on the validity of measures of subjective well-being**

Type of evidence	Sources
Face validity	
• Item-specific non-response rates. • Time to reply.	Rässler and Riphahn (2006); Smith (2013); ONS (2011);
Convergent validity	
• Ratings by friends and family.	Frey and Stutzer (2002); Pavot and Diener (1993); Schneider and Schimmack (2009).
• Ratings by interviewers.	Pavot and Diener (1993).
• Emotion judgements by strangers.	Diener, Suh, Lucas and Smith (1999).
• Frequency/intensity of smiling.	Frey and Stutzer (2002); Kahneman and Krueger (2006); Seder and Oishi (2012).
• Changes in behaviour.	Frijters (2000); Diener (2011); Clark, Georgellis and Sanfrey (1998).
• Biophysical measures.	Urry et al. (2004); Steptoe, Wardle and Marmot (2005); Kahneman and Krueger (2006)
• Relationships among different evaluative, affective and/or eudaimonic measures.	Diener, Helliwell and Kahneman (2010); Kahneman and Krueger (2006), Clark and Senik (2011); Diener, Wirtz, Biswas-Diener, Tov, Kim-Prieto, Choi and Oishi (2009); Huppert and So (2009)
Construct validity	
• Association with income (individual and national level).	Sacks, Stevenson and Wolfers (2010).
• Life events (e.g. impact of becoming unemployed, married, disabled, divorced or widowed).	Diener, Lucas and Napa Scollon (2006); Lucas (2007); Lucas, Clark, Georgellis and Diener (2003); Winkelmann and Winkelmann (1998).
• Life circumstances (health status, education, social contact, being in a stable relationship).	Dolan, Peasgood and White (2008); NEF (2009).
• Daily activities (e.g. commuting, socialising, relaxing, eating, praying, working, childcare, housework).	Kahneman and Krueger (2006); Frey and Stutzer (2008); Helliwell and Wang (2011); Stone (2011).

Subjective well-being measures can also be sensitive to specific aspects of the survey content. For example, Deaton (2011) found that asking questions about whether or not the country is going in the right direction immediately before an evaluative subjective well-being measure exerted a strong downward influence on the data. Similarly, a number of authors have shown a question order effect when life satisfaction and dating or marriage satisfaction questions are asked (e.g. Strack, Martin and Schwarz, 1988; Schwarz, Strack and Mai, 1991; Tourangeau, Rasinski and Bradburn, 1991). Pudney (2010) finds some evidence that the survey mode impacts the relationship between different satisfaction domains and their objective drivers. These effects are real and can have significant implications for measurement and fitness for use. However, they are largely factors that have the potential to be managed through consistent survey design. Chapter 2 discusses these issues and the best way of handling them.

Differences may also exist among respondents in terms of how questions are interpreted, how different response formats and scale numbers are used, and the existence of certain response styles, such as extreme responding and acquiescence. Socially desirable responding may also impact the mean levels of reported subjective well-being. The evidence for these effects, and the methodological implications, are discussed in further detail in Chapter 2. To the extent that these differences are randomly distributed among populations of interest, they will contribute to random "noise" in the data without necessarily posing a fundamental challenge to data users. However, where they systematically vary across different survey methods, and/or where they affect certain groups, nationalities or cultures differently, this can make the interpretation of group and sample differences in subjective well-being problematic.

Group differences in scale use may arise for a number of reasons, including translation issues (e.g. Veenhoven, 2008; Oishi, 2010), differences in susceptibility to certain response styles (Hamamura, Heine and Paulhus, 2008; Minkov, 2009; van Herk, Poortinga and Verhallen, 2004), or the cultural relevance and sensitivity of certain subjective well-being questions (Oishi, 2006; Vittersø, Biswas-Diener and Diener, 2005). Various methods do exist to detect and control for these effects, and survey design can also seek to minimise this variability, as is described in Chapter 2. However, further research is needed to inform the best approach to international comparisons in particular. That said, there is evidence to suggest that these responses do not extensively bias the analysis of determinants in micro-data (Helliwell, 2008; Fleche, Smith and Sorsa, 2011). At the national level, there remain clear and consistent relationships between objective life circumstances and subjective well-being (Helliwell, 2008), and these differences are reflected in mean scores – such that, for example, the distribution of life satisfaction scores in Denmark and Togo are almost non-overlapping (Diener, 2011).

A final consideration is the extent to which individual, cultural or national fixed effects influence subjective well-being measures – including differences in personality and dispositional affect (Diener, Oishi and Lucas, 2003; Diener, Suh, Lucas and Smith, 1999; Schimmack, Diener and Oishi, 2002; Suls and Martin, 2005). These differences may be due to genetic factors (e.g. Lykken and Tellegen, 1996) or environmental factors in a person's development that produce chronically accessible and stable sources of information on which subjective well-being assessments may be based (e.g. Schimmack, Oishi and Diener, 2002). To the extent that public policy can shape a person's life experiences, particularly in developmental phases, the existence of large and relatively stable individual fixed effects does not necessarily mean that measures are insensitive to the effects of policy interventions – but the time frames over which these experiences take effect may be quite long.

It is clear that individual fixed effects are real, and they account for a significant proportion of variance across individuals. For example, in two national longitudinal panel studies, Lucas and Donnellan (2011) found that 34-38% of variance in life satisfaction was due to stable trait differences, and a further 29-34% of additional variance was due to an autoregressive component that was moderately stable in the short-term, but could fluctuate over longer time periods. However, this study did not include measures of the objective life circumstances that might impact on both stable trait-like and autoregressive components. Different cultures and nations also vary in both mean levels (e.g. Bjørnskov, 2010; OECD, 2011) and in the composition or construction of reported subjective well-being (e.g. Schimmack, Oishi and Diener, 2002; Diener, Napa Scollon, Oishi, Dzokoto and Suh, 2000; Oishi, 2006) – although again, it is rare to find research that explicitly documents the contribution of national or cultural fixed effects over and above objective life circumstances.

It is possible that individual, cultural and national fixed effects are substantive, in so far as they have a genuine impact on how people subjectively feel about their well-being. If this is the case, they should not be regarded as measurement error. In practice, however, it is not always easy to disentangle fixed effects in actual experienced subjective well-being from differences in translation, response styles, question meaning and retrospective recall biases – although Oishi (2010) suggests that measurement artefacts (such as differences in number use, item functioning and self-presentation) play a relatively small role in explaining overall national differences at the mean level. What we do know is that, although there is a reasonably stable component to life evaluations, these measures are still sensitive to life circumstances, and do change over time in response to life events (Lucas, 2007; Lucas, Clark, Georgellis and Diener, 2003; Diener, Lucas and Napa Scollon,

2006). Thus, the main drivers of subjective well-being of interest to policy-makers, including long-term life circumstances that may influence resilience to the impact of negative life events, can still be examined. Panel data can also be used so that the impact of changes in life circumstances (including policy interventions) can be examined whilst controlling for fixed effects where necessary.

Summary: Limits of validity

While there are a wide range of issues that potentially limit the validity of subjective measures of well-being, many of these are either of marginal impact in terms of fitness for purpose (i.e. they do not substantively affect the conclusions reached) or can be dealt with through appropriate survey design. For example, any contextual effects from the preceding question will not bias the analysis of changes over time if the survey itself does not change over time. In practical terms, the impact of these limitations on fitness for purpose depends very much on the purpose for which the data is being used. These issues will be dealt with in Chapter 2. One major limit does, however, need to be acknowledged. Despite evidence that cultural factors do not substantively bias multi-variate analysis, there is good reason to be cautious about cross-country comparisons of levels of subjective well-being – particularly life satisfaction.

Coherence and the measurement of subjective well-being

Coherence addresses the degree to which different measures of the same underlying construct tell the same story. Two similar statistics might reflect slightly different concepts, and hence not be comparable even though they are both highly accurate. Coherence depends on the use of common concepts, definitions and classifications. For this reason, the issue of coherence in measures of subjective well-being essentially reduces to the case for a common measurement framework. As discussed in the first section of this chapter, this is the primary rationale for producing a set of guidelines. While the guidelines themselves will not initially constitute a standard, they are a necessary initial step in the process that might later lead to a formal standard.

Conclusion

The case for national statistical agencies producing measures of subjective well-being depends on the relevancy and accuracy of such measures. If it is possible to produce measures of subjective well-being that are of sufficient quality to be fit for purpose, and which are of use either in formulating policy or in informing the broader public in useful ways, then there is a strong argument in favour of collecting and publishing such measures. Over the last ten years, measures of subjective well-being have moved increasingly from the realm of academic discourse into the realm of informing policy-making and public debate. There is now a sufficient body of knowledge on how measures of subjective well-being will be used to make a strong *prima facie* case for their relevance. While ultimately the proof of relevance will depend on up-take and use, this cannot occur until measures are produced, at least on an experimental basis. For this reason it is important that at least some official statistical agencies start producing regular series of subjective well-being data to support policy-makers and allow more informed decisions about the ultimate usefulness of such data. Only when such data is available for some time can more definite judgements be reached on relevance.

There is a large body of evidence on the reliability and validity of measures of subjective well-being and the methodological challenges involved in collecting and analysing such data. Indeed, given the academic interest in the topic and the challenging nature of the subject, the body of evidence on the strengths and weaknesses of measures of subjective well-being exceeds that for many measures regularly collected as part of official statistics (Smith, 2013). While measures of subjective well-being have some important limitations, there is no case for simply considering subjective measures of well-being "beyond the scope" of official statistics. Although subject to some methodological limitations, it is clear that for many of the purposes for which they are desired, measures of subjective well-being are able to meet the basic standard of "fit for purpose".

However, there are also a number of areas where measures of subjective well-being are more strongly affected by issues to do with how the measure is collected and with potentially irrelevant characteristics of the respondent than is the case for many other official statistics. This does not suggest that measures of subjective well-being should be rejected outright, but points to two important steps. First, there is a need for official measures of subjective well-being to be collected – possibly as experimental data series – in order to provide the basis to resolve some of the outstanding methodological issues. Second, it is important that information on the nature of the most significant methodological issues associated with collecting measures of subjective well-being is available to potential producers of subjective well-being data, and that a common approach to dealing with these issues is developed. This is the focus of the second chapter.

Notes

1. E.g. "A New Gauge to See What's Beyond Happiness", *New York Times*, 16 May 2011.
2. The definition used is taken largely from Diener et al. (2006).
3. By "decision utility" Kahneman refers to the sort of evaluation used by individuals to make choices between different options. He distinguishes this from "experienced utility" which is the sum of felicific experience for an individual over time. The former approaches what economists mean by utility in standard microeconomic models, while the latter is closer to Jeremy Bentham's original notion of utility in the context of utilitarianism (Bentham, 1789).
4. For example, the Gallup World Poll contains a range of questions on affect during the previous day, which have been extensively tested. The UK Office of National Statistics has collected similar measures of affect in its Integrated Household Survey programme.
5. One concern sometimes raised about subjective measures is that they are unlikely to change as fast over time as more traditional indicators. In fact, this is not strictly true (see Box 1.2), However, even for those circumstances where measures of subjective well-being do not change as much as, say, resource-based measures, this should be regarded as information rather than as a problem with the measure.
6. There are two primary reasons why subjective well-being might be considered to differ substantially from overall well-being. First, subjective well-being is affected by a number of factors, such as personality and culture, which might be considered a source of bias in terms of measuring actual well-being. Second, most theories of well-being are not strictly utilitarian in nature and recognise standards that are important to well-being regardless of their association with a person's subjective state. For example, Sen (1979) defines well-being in terms of achieved "capabilities" to do certain things, explicitly rejecting a subjective (utilitarian) alternative.
7. Consideration of initial sample variance in each measure is important here: if the sample has uniformly high levels of health satisfaction, but variable levels of housing satisfaction, housing satisfaction may look more important in a regression analysis, simply because it has more variation to associate with variation in the outcome measure.

8. Cronbach's coefficient alpha is considered to be the most stable and accurate index of internal consistency reliability (Kline, 2000; Nunnally and Bernstein, 1994). Provided all item standard deviations are equal, alpha is mathematically equivalent to the mean of all split-half reliabilities (i.e. dividing test items into two halves and correlating those halves with one another); it is slightly lower than the mean where item standard deviations vary (Cortina, 1993). Alpha is calculated by multiplying the mean average inter-item co-variance by the number of items in a test (which estimates the true score variance) and dividing this figure by the sum of all the elements in the variance-covariance matrix, which equals the observed test variance (Nunnally and Bernstein, 1994).
9. Not all of this "noise" is strictly error, however. The sensitivity of affect measures to the day of the week, for example, validates these measures to the extent that individuals participate in more pleasurable activities, such as time spent with friends and family, on weekends (Heliwell and Wang, 2011; Stone, 2011).

Bibliography

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/*.

Bartam, D. (2010), "Economic Migration and Happiness: Comparing Immigrants' and Natives' Happiness Gains From Income", *Social Indicators Research*, Vol. 103, No. 1, pp. 57-76.

Bentham, J. (1789), *An Introduction to the Principles of Morals and Legislation*, Oxford, Clarendon Press.

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

Boarini, R., M. Comola, C. Smith, R. Manchin and F. De Keulenaer (2012), "What makes for a better life: the determinants of subjective well-being in OECD countries, evidence from the Gallup World Poll", *OECD STD Working Paper*.

Carrin, G., D. Evans, F. Ravndal and K. Xu (2009), "Assessing the reliability of household expenditure data: Results of the World Health Survey", *Health Policy*, Vol. 91(3), pp. 297-305.

Clark, A. (2001), "What really matters in a job? Hedonic measurement using quit data", *Labour Economics*, No. 8, pp. 223-242.

Clark, A.E., Y. Georgellis and P. Sanfey (1998), "Job Satisfaction, Wage Changes and Quits: Evidence from Germany", *Research in Labor Economics*, Vol. 17.

Clark, A. and A. Oswald (2002), "A simple statistical method for measuring how life events affect happiness", *International Journal of Epidemiology*, No. 31, pp. 1139-1144.

Clark, A. and C. Senik (2011), "Is happiness different from flourishing? Cross-country evidence from the ESS", *Working Paper 2011-04*, Paris, School of Economics.

Clifton, J. and L. Morales (2011), "Egyptians', Tunisians' Wellbeing Plummets Despite GDP Gains", Gallup Organisation.

Cortina, J.M. (1993), "What is coefficient alpha? An examination of theory and applications", *Journal of Applied Psychology*, Vol. 78(1), pp. 98-104.

Cummins, R., R. Eckersley, J. Pallant, J. Vugt and R. Misajon (2003), "Developing a National Index of Subjective Wellbeing: The Australian Unity Wellbeing Index", *Social Indicators Research*, Vol. 64, No. 2, pp. 159-190.

Deaton, A. (2011), "The financial crisis and the well-being of Americans", *Oxford Economic Papers*, No. 64, pp. 1-26.

Deci, E. and R. Ryan (2006), "Hedonia, Eudaimonia, and Well-being: An introduction", *Journal of Happiness Studies*, No. 9, pp. 1-11.

Di Tella, R., R. MacCulloch and A. Oswald (2003), "The Macroeconomics of Happiness", *The Review of Economics and Statistics*, Vol. 85(4), pp. 809-827.

Diener, E. (2011), *The Validity of Life Satisfaction Measures*, unpublished working paper.

Diener, E. (2006), "Guidelines for National Indicators of Subjective Well-Being and Ill-Being", *Applied Research in Quality of Life*, No. 1, pp. 151-157.

Diener, E. (1984), "Subjective Well-Being", *Psychological Bulletin*, Vol. 95, No. 3, pp. 542-575.

Diener, E. and M. Chan (2011), "Happy People Live Longer: Subjective Well-being Contributes to Health and Longevity", *Applied Psychology: Health and Well-being*, March.

Diener, E., R. Emmons, R. Larsen and S. Griffin (1985), "The Satisfaction With Life Scale", *Journal of Personality Assessment*, Vol. 49(1), pp. 71-75.

Diener, E., J.F. Helliwell and D. Kahneman (eds.) (2010), *International Differences in Well-Being*, Oxford University Press.

Diener, E., R.E. Lucas, U. Schimmack and J. Helliwell (eds.) (2009), *Well-Being For Public Policy*, Oxford University Press.

Diener, E., R.E. Lucas and C. Napa Scollon (2006), "Beyond the hedonic treadmill: Revising the adaptation theory of well-being", *American Psychologist*, Vol. 61(4), pp. 305-314.

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., C.K.N. Scollon, S. Oishi, V. Dzokoto and E.M. Suh (2000), "Positivity and the construction of life satisfaction judgements, global happiness is not the sum of its parts", *Journal of Happiness Studies*, No. 1, pp. 159-176.

Diener, E., E. Suh, R. Lucas and H. 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. Choi and S. Oishi (2009), "New Measures of Well-Being", in E. Diener (ed.) (2009), *Assessing Well-Being: The Collected Works of Ed Diener*, Social Indicators Research Series, Vol. 39, Dordrecht: Springer, pp. 247-266.

Dolan, P., R. Layard and R. Metcalfe (2011), *Measuring Subjective Well-being for Public Policy*, Office for National Statistics.

Dolan, P. and T. Peasgood (2006), "Valuing non-market goods: Does subjective well-being offer a viable alternative to contingent valuation", *Imperial College Working Paper*, Imperial College, London.

Dolan P., T. Peasgood and M. White (2008), "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being", *Journal of Economic Psychology*, Vol. 29, pp. 94-122.

Dolan, P. and M. White (2007), "How Can Measures of Subjective Well-Being Be Used to Inform Public Policy?", *Perspectives on Psychological Science*, Vol. 2(1), pp. 71-85.

Easterlin, R. (1974), "Does Economic Growth Improve the Human Lot? Some Empirical Evidence", in David, P.A. and M.W. Reder, *Nations and Households in Economic Growth: Essays in Honour of Moses Abramovitz*, New York, Academic Press Inc, pp. 89-125.

Ferrer-i-Carbonell, A. and P. Frijters (2004), "How important is methodology for the estimates of the determinants of happiness?", *The Economic Journal*, No. 114, pp. 641-659.

Fleche, S., C. Smith and P. Sorsa (2011), "Exploring determinants of subjective wellbeing in OECD countries-evidence from the World Value Survey", *OECD Working Paper*, No. 921.

Frey, B.S. and A. Stutzer (2011), "The use of happiness research for public policy", *Social Choice Welfare*, Vol. 38(4), pp. 659-674.

Frey, B.S. and A. Stutzer (2008), "Stress that Doesn't Pay: The Commuting Paradox", *The Scandinavian Journal of Economics*, Vol. 110(2), pp. 339-366.

Frey, B.S. and A. Stutzer (2005), "Happiness Research: State and Prospects", *Review of Social Economy*, Vol. 63(2), pp. 207-228.

Frey, B.S. and A. Stutzer (2002), "What Can Economists Learn from Happiness Research?", *Journal of Economic Literature*, Vol. 40(2), pp. 402-435.

Frijters, P. (2000), "Do individuals try to maximize general satisfaction?", *Journal of Economic Psychology*, No. 21, pp. 281-304.

Fujita, F. and E. Diener (2005), "Life satisfaction set point: stability and change", *Journal of Personality and Social Psychology*, Vol. 88(1), p. 158.

Fujiwara, D. and R. Campbell (2011), "Valuation Techniques for Social Cost-Benefit Analysis: Stated Preference, Revealed Preference and Subjective Well-Being Approaches", *Green Book Discussion Paper*.

Gallup (2011), "Egyptians', Tunisians' Wellbeing Plummets Despite GDP Gains", available online at: *www.gallup.com/poll/145883/egyptians-tunisians-wellbeing-plummets-despite-gdp-gains.aspx*.

Halpern, D. (2010), *The Hidden Wealth of Nations*, Polity Press.

Hamamura, T., S.J. Heine and D.L. Paulhus (2008), "Cultural differences in response styles: The role of dialectical thinking", *Personality and Individual Differences*, No. 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.

Helliwell, J.F. (2008), "Life Satisfaction and the Quality of Development", *NBER Working Paper*, No. 14507, National Bureau of Economic Research.

Helliwell, J.F. and C.P. Barrington-Leigh (2010), "Measuring and Understanding Subjective Well-being", *NBER Working Paper*, No. 15887, National Bureau of Economic Research.

Helliwell, J. and S. Wang (2011), "Trust and Well-being", *International Journal of Well-being*, Vol. 1(1), pp. 42-78.

Hicks, S. and L. Tinkler (2011), "Measuring Subjective Well-being", *ONS Supplementary Paper*.

HM Treasury, (2011), *The Green Book: Appraisal and Evaluation in Central Government*.

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 NSOs?*, Florence, 23/24 July, available online at: *www.isqols2009.istitutodeglinnocenti.it/Content_en/Huppert.pdf*.

Huppert, F.A., N. Marks, A. Clark, J. Siegrist, A. Stutzer, J. Vitterso and M. Wahrendorf (2009), "Measuring Well-being Across Europe: Description of the ESS Well-being Module and Preliminary Findings", *Social Indicators Research*, No. 91, pp. 301-315.

International Wellbeing Group (2006), *Personal Wellbeing Index*, 4th edition, Melbourne, Australian Centre on Quality of Life, Deakin University, available online at: *www.deakin.edu.au/research/acqol/instruments/wellbeing_index.htm*.

Kahneman, D. and A. Deaton (2010), "High income improves life evaluation but not emotional well-being", *Proceedings of the National Academy of Sciences*, Vol. 107(38), pp. 16489-16493.

Kahneman, D., E. Diener and N. Schwarz (1999), *Well-being: The Foundations of Hedonic Psychology*, Russel Sage Foundation, New York.

Kahneman, D. and A. Krueger (2006), "Developments in the measurement of subjective well-being", *The Journal of Economic Perspectives*, No. 20, pp. 3-24.

Kahneman, D., A. Krueger, D. Schkade, N. Schwarz and A. Stone (2004), "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method", *Science*, Vol. 306(5702), pp. 1776-1780.

Kline, P. (2000), *The Handbook of Psychological Testing*, 2nd edition, Florence, KY, US: Taylor and Frances/Routledge.

Krueger, A. and D. Schkade (2008), "The reliability of subjective well-being measures", *Journal of Public Economics*, No. 92, pp. 1833-1845.

Krueger, A., D. Kahneman, C. Fischler, D. Schkade, N. Schwarz and A. Stone (2009), "Time Use and Subjective Well-Being in France and the US", *Social Indicators Research*, No. 93, pp. 7-18.

Krueger, A., D. Kahneman, D. Schkade, N. Schwarz and A. Stone (2009), "National Time Accounting: The Currency of Life", in A. Krueger (ed.), *Measuring the Subjective Well-being of Nations: National Accounts of Time Use and Well-Being*, NBER.

Larson, R.J. and B.L. Fredrickson (1999), "Measurement Issues in Emotion Research", in D. Kahneman, E. Diener and N. Schwarz (eds.), *Well-being. The Foundations of Hedonic Psychology*, Russel Sage Foundation, New York, pp. 40-60.

Lucas, R.E. (2007), "Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies", *Journal of Personality and Social Psychology*, Vol. 92(4), pp. 717-730.

Lucas, R.E., A.E. Clark, Y. Georgellis and E. Diener (2003), "Re-examining adaptation and the set point model of happiness: Reactions to changes in marital status", *Journal of Personality and Social Psychology*, Vol. 84, pp. 527-539.

Lucas, R.E. and M. Donnellan (2012), "Estimating the Reliability of Single-Item Life Satisfaction Measures: Results from Four National Panel Studies", *Social Indicators Research*.

Lykken, D. and A. Tellegen (1996), "Happiness is a stochastic phenomenon", *Psychological Science*, No. 7, pp. 186-189.

Michalos, A. and P. Kahlke (2010), "Stability and Sensitivity in Perceived Quality of Life Measures: Some Panel Results", *Social Indicators Research*, No. 98, pp. 403-434.

Minkov, M. (2009), "Nations with more dialectical selves exhibit lower polarization in life quality judgements and social opinions", *Cross-Cultural Research*, No. 43, pp. 230-250.

New Economics Foundation, (2009), *National Accounts of Well-being*.

Nunnally, J.C. and I.H. Bernstein (1994), *Psychometric Theory (Third Edition)*, New York: McGraw-Hill.

OECD (2011), *How's Life? Measuring Well-being*, OECD Publishing.

OECD (2008), *Quality Framework and Guidelines for OECD Statistical Activities*, OECD Publishing.

Oishi, S. (2010), "Culture and well-being: Conceptual and methodological issues", in E. Diener, J.F. Helliwell and D. Kahneman (eds.), *International Differences in Well-Being*, Oxford: Oxford University Press.

Oishi, S. (2006), "The concept of life satisfaction across cultures: An IRT analysis", *Journal of Research in Personality*, Vol. 40(4), pp. 411-423.

ONS (2012), *Subjective Well-being User Guide: 12 Month Data Set*, Office for National Statistics.

ONS (2011a), *Initial investigation into Subjective Wellbeing from the Opinions Survey*, Office for National Statistics.

ONS (2011b), *Measuring Subjective Well-being*, Office for National Statistics, United Kingdom.

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

Pavot, W., E. Diener, C.R. Colvin and E. Sandvik (1991), "Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures", *Journal of Personality Assessment*, Vol. 57(1), pp. 149-161.

Pudney, S. (2010), *An experimental analysis of the impact of survey design on measures and models of subjective wellbeing*, Institute for Social and Economic Research.

Rässler, S. and R.T. Riphahn (2006), "Survey item nonresponse and its treatment", *Allgemeines Statistisches Archiv*, No. 90, pp. 217-232.

Sacks, W.D., B. Stevenson and J. Wolfers (2010), "Subjective Well-being, Income, Economic Development and Growth", *NBER Working Paper*, No. 16441.

Schimmack, U., E. Diener and S. Oishi (2002), "Life satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources", *Journal of Personality*, Vol. 70(3), pp. 345-382.

Schimmack, U., S. Oishi and E. Diener (2002), "Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivisim?", *Cognition and Emotion*, Vol. 16(6), pp. 705-719.

Schneider, L. and U. Schimmack (2009), "Self-Informant Agreement in Well-Being Ratings: A Meta-Analysis", *Social Indicators Research*, Vol. 94(3), pp. 363-376.

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.

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.

Seder, J.P. and S. Oishi (2012), "Intensity of smiling in facebook photos predicts future life satisfaction", *Social Psychological and Personality Science*, Vol. 3(4), pp. 407-413.

Sen, A. (1979), *Equality of What?*, Tanner Lecture on Human Values, Stanford University.

Smith, C. (2013), "Making happiness count: Four myths about subjective well-being", *OECD Working Paper*, forthcoming.

Steptoe, A., J. Wardle and M. Marmot (2005). "Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102(18), pp. 6508-6512.

Stevenson, W. (2008), "Economic Growth and Subjective Wellbeing: Reassessing the Easterlin Paradox", *NBER Working Paper*, No. 14282.

Stiglitz, J.E., A. Sen and J.P. Fitoussi (2008), *Report by the Commission on the Measurement of Economic Performance and Social Progress*.

Stone, A.A. (2011) "A rationale for including a brief assessment of hedonic well-being in large-scale surveys", *Forum for Health Economics and Policy*, Vol. 14(3), Article 7.

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.

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(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.

Thompson, S. and N. Marks (2008), *Measuring well-being in policy: issues and applications*, New Economics Foundation.

Tourangeau, R. (1999), "Context Effects on Answers to Attitude Questions", in G. Monroe (ed.), *Cognition and Survey Research*, Sirken, Wiley, New York, pp. 111-131.

Tourangeau, R., K.A. Rasinski and N. Bradburn (1991), "Measuring Happiness in Surveys: A Test of the Subtraction Hypothesis", *Public Opinion Quarterly*, No. 55, pp. 255-266.

Urry, H., J. Nitschke, I. Dolski, D. Jackson, K. Dalton, C. Mueller, M. Rosenkranz, C. Ruff, B. Singer and R. Davidson (2004), "Making a Life Worth Living: Neural Correlates of Well-Being", *Psychological Science*, Vol. 15, No. 6, pp. 367-372.

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*, No. 35, pp. 346-360.

Van Praag, B.M.S., P. Frijters and A. Ferrer-i-Carbonell (2003), "The anatomy of subjective well-being", *Journal of Economic Behaviour and Organisation*, No. 51, pp. 29-49.

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.

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(6), pp. 1063-1070.

Weinhold, D. (2008), "How big a problem is noise pollution? A brief happiness analysis by a perturbable economist", *MPRA Working Paper*, No. 10660.

Wilson, T.D. and D.T. Gilbert (2006), "Affective forecasting: Knowing what to want", *Current Directions in Psychological Science*, Vol. 14(3), pp. 131-134.

Wilson, T.D., T. Wheatley, J.M. Meyers, D.T. Gilbert and D. Axsom (2000), "Focalism: A source of durability bias in affective forecasting", *Journal of Personality and Social Psychology*, Vol. 78(5), pp. 821-836.

Winkelman, L. and R. Winkelman (1998), "Why Are the Unemployed So Unhappy? Evidence from Panel Data?", *Economica*, Vol. 65, pp. 1-15.

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.

Chapter 3

Measuring subjective well-being

Introduction

This chapter aims to present best practice in measuring subjective well-being. It covers both the range of concepts to be measured and the best approaches for measuring them. This includes considering issues of sample design, survey design, data processing and coding and questionnaire design. In particular, the chapter presents a single primary measure intended to be collected consistently across countries, as well as a small group of core measures that it is desirable for data producers to collect where possible. Beyond this core suite of measures, the chapter provides more general advice to support data producers interested in identifying and measuring aspects of subjective well-being that will meet their particular research or policy needs, as well as a range of question modules relating to different aspects of subjective well-being.

The chapter has four substantive sections. The first section focuses on issues associated with planning the measurement of subjective well-being. This includes addressing the relationship between the intended policy or research use of the data and the appropriate measurement objectives. A crucial element of deciding what to measure is thinking about the relevant co-variates to be collected alongside the measures of subjective well-being to support analysis and interpretation. Section 2 of the chapter addresses survey and sample design issues. These include the choice of survey vehicle, sample design, target population, collection period and survey frequency. The third section of the chapter looks at questionnaire design, which includes both issues of question order and questionnaire structure, as well as the precise question wording. A key element of this section is the inclusion of model questions on the different aspects of subjective well-being. The final section focuses on survey implementation. This includes brief guidelines on interviewer training as well as data processing. The chapter does not, however, cover issues relating to the use and analysis of subjective well-being data in detail. These are addressed in Chapter 4 (*Output and analysis of subjective well-being measures*). A recurring issue throughout the chapter is the lack of standards in the methods used to gather supporting information for subjective well-being analyses, such as information about child well-being, attitudes, personality, etc. These issues are deemed to be beyond the scope of this chapter, but remain an important gap.

Core measures of subjective well-being

Core measures of subjective well-being are those for which international comparability is the highest priority. These are measures for which there is the most evidence of validity and relevance, for which the results are best understood, and for which policy uses are most developed. Although the guidelines are intended to support producers of measures of subjective well-being rather than being overly prescriptive, the core measures proposed here are quite specific in content and collection method.

The core measures outlined in this chapter consist of five questions. The first is a primary measure and is intended to be collected consistently across countries. The primary measure should be regarded as the highest priority for national statistical agencies and should be the first question included in surveys where the measurement of subjective well-being is considered. The additional three affect questions in the core are also important and should be collected where possible. However, it is recognised that not all national statistical offices will be able to collect these measures in their core surveys. Finally, an experimental eudaimonic question is attached, picking up the element of eudaimonia for which there is the most evidence of relevance.

Beyond articulating a suite of core measures, the main goal of this chapter is to provide general advice to data providers. In particular, the chapter is intended to support national statistical agencies and other data providers in the process of deciding what to measure and how to implement the measurement process most effectively. While models are provided for specific questions, the chapter aims to provide options and advice rather than directions.

1. What to measure? Planning the measurement of subjective well-being

This section looks at the planning stage of a measurement project. It is concerned with what concepts to measure and how these concepts affect decisions about the final output and analysis. In doing so, the chapter touches on the issues that are the main focus of Chapter 4 (*Output and analysis of subjective well-being measures*). However, where Chapter 4 focuses on how to analyse, interpret and present subjective well-being data, the discussion here is limited to how user needs determine what information to collect.

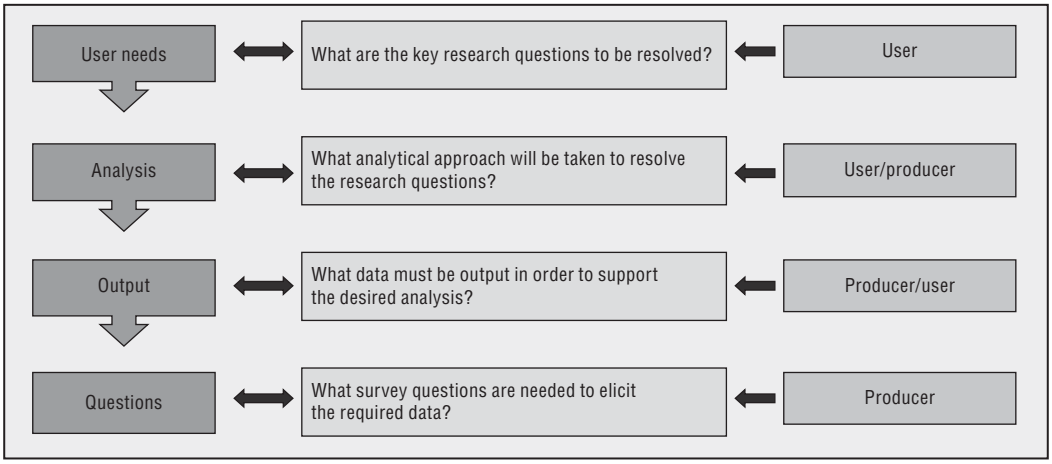
The initial planning stage of a project to measure subjective well-being – or indeed any statistical programme – is critically important. All subsequent decisions will be heavily influenced by choices made early on about the research objectives of the project. Clarity about objectives is thus crucial.

Decisions about what to measure should always be grounded in a clear understanding of user needs. Only if the needs for the data are clearly understood is it possible to make informed decisions about the information that should be collected to meet these needs. Understanding user needs is not, however, straight-forward. A relatively simple research question can be approached in a range of different ways using different methodologies. For example, one can understand what motivates behaviour both by asking people directly what they would do in a given set of circumstances or by collecting information on the course of action people take and on the circumstances they face.[1] Each methodological approach has its own strengths and weaknesses, and will have different implications for measurement. Having an analytical model can assist in thinking in a structured way about how user needs relate to specific decisions about what data to collect.

Figure 3.1 presents a simple model relating user needs to the specific survey questions used to collect information. The model is intended to provide a framework for thinking about the various stages involved in moving from a user's information needs to specific questions that can be included in a survey.

The first column of Figure 3.1 identifies the four stages involved in going from user needs to specific survey content. Conceptually, these stages involve working back through the process of collecting the data and using them in decision-making in reverse order. Column 2 articulates the key issues to be addressed in each stage of the project in order to make well-informed decisions about the most appropriate measures. Finally, the third

Figure 3.1. **The planning process: From user needs to survey questions**



column indicates which party has the lead role in making decisions. Although the process of going from user needs to survey content is fundamentally collaborative in nature, there are stages in the process when users can be expected to play a more important role than data providers, as well as cases where the reverse is true.

In practice, the process of working through these four stages is likely to be less clearly defined than Figure 3.1 suggests. In some cases, where the level of analysis required is relatively simple, the analysis and output stages of the process can merge into each other. Users will sometimes have clear views about the best measures to support the analysis that they would like to undertake, and it would be foolish to ignore these views in instances where a sophisticated user has a better understanding of the issue at hand than a data provider with little experience of measuring subjective well-being. Similarly, data producers may suggest possibilities that will result in changes in user needs or in the analytical approach taken to address them.

User needs

Understanding user needs involves understanding the key policy and research questions that the user is trying to address. While it is not possible in this chapter to give a full discussion of all possible user needs for subjective well-being data, some general questions can be articulated:

- Are the user needs related to one of the general policy uses for subjective well-being data described in Chapter 1?
- What are the policy questions?
- Is the subjective wellbeing content being proposed appropriate to respond to the policy questions? Is the content proposed sensitive to monitoring changes over time or between population groups?
- What population groups are of interest to the user? For example, is the focus on international comparisons (making countries the key unit of analysis), the same population at different points in time (for time series analysis), or different sub-groups of the same population (such as age, sex, location or ethnicity)? This will have implications both for sampling and for the types of measure that are most appropriate.

In the case of cross-country comparisons, measures with good cross-cultural reliability will be most important, while for analysis of groups within a country low respondent burden may be a more important consideration in order to allow a larger sample size.

- Does the user's interest lie in comparing outcomes of different groups or in understanding the relationship between different aspects of subjective well-being? In the first case, a relatively narrow range of subjective well-being measures may suffice, while in the latter case more detail on a range of co-variates is likely to be necessary.
- Is the user's primary interest in overall subjective well-being (captured by summary measures of life evaluation, affect or eudaimonic well-being) or in a specific dimension of subjective well-being (such as satisfaction with income or satisfaction with work/life balance)? Are other measures of well-being more appropriate?
- What are the frequency requirements of the users to monitor changes over time?
- What within-country comparisons are required, such as geographic level?

A thorough understanding of user needs should allow the identification of one or more clear research questions that the project should address.

Analysis

Understanding the overall research question is not sufficient to make meaningful decisions about the type of output or the most appropriate measures to use. A given research question may be addressed in more than one valid way. It is therefore essential to understand how the specific research question can be answered:

- Will the analytical approach be primarily descriptive, or will it require more sophisticated statistical techniques (e.g. regression, factor analysis, etc.)?
- What contextual and other variables are required to answer the research question? If the research question simply involves identifying differences between specific population groups in terms of a small set of key outcomes, the range of relevant co-variates may be relatively limited. However, if the research question is focused on understanding what drives group differences in subjective well-being or on examining the joint distribution between subjective well-being and other dimensions of well-being, the range of co-variates is likely to be significantly broader.
- What level of accuracy is required to produce meaningful results from the proposed analysis? This will have implications for sample size and sampling strategy. For example, if obtaining precise estimates for small population sub-groups is a priority, then oversampling of these groups may be necessary.

After considering the proposed analytical strategy, it should be possible to articulate how the research questions can be answered in quite specific terms. This will form the basis for evaluating what data needs to be output to support the desired analysis.

Output

Output refers to the statistical measures released by a national statistical agency or by another data producer. These can take the form of tables of aggregate data such as average results by group, micro-data files, interactive data cubes or other forms. The key distinction between output and analysis is that output does not, in itself, answer a research question. Instead, it provides the base information that is analysed in order to produce the answer.

In some cases, the answer may be directly evident from the output, requiring only limited interpretation, comment and caveats, while in other cases extensive statistical analysis may be required.

Because output forms the basis for all subsequent analysis, it provides the key link between specific survey questions and the use of the data in analysis. The required output must therefore be clearly specified before appropriate questions can be designed. Some key issues to consider when specifying the desired output for information on subjective well-being include:

- Will the analysis require tabular output of averages or proportions, or is micro-data needed? Simple comparisons of how different population groups compare with each other can be accomplished via tabular output, but understanding the drivers of such differences will require a much finer level of detail.
- Will the analytic techniques used treat the data as ordinal or cardinal? This makes little difference if micro-data is required (since users can decide for themselves), but will influence how summary measures of central tendency and distribution are presented in tabular form. Information on a cardinal variable[2] can be presented via techniques that add and average scores (e.g. mean, standard deviation), while ordinal data will need to be reported by category.
- How important is it to present measures of the central tendency of the data (e.g. mean, median, mode) as opposed to the dispersal (e.g. standard deviation) or full distribution of the data (e.g. proportion responding by category)?

In planning a measurement exercise, the aim should be to clearly specify the desired output, and the data items required to produce this, before considering question design. This will involve, at a minimum, defining the measures to be used and the break-downs and cross-classifications required. In many cases, particularly if multivariate analysis is proposed, more detailed information may be required.

Questionnaire design

Once a clear set of outputs has been identified based on the analysis required to meet user needs, it will be possible to make specific decisions about survey design, including the most appropriate survey vehicle, collection period, units of measurement and questionnaire design. These decisions should flow logically from the process of working down from user needs through analysis and output. The remainder of this chapter sets out a strategy for the measurement of subjective well-being. This includes both specific proposals for how a national statistical agency might approach the measurement of subjective well-being and more general information that can be used in a wider range of circumstances.

What other information should be collected: Co-variates and analytical variables

All potential uses of subjective well-being data require some understanding of how subjective well-being varies with respect to other variables. This applies whether the goal is understanding the drivers of subjective well-being – which requires understanding the causes of change – or where the main purpose is monitoring well-being over time and across countries – which requires understanding changes in demographics in order to understand a given change is due to changes in average levels or in the ratios of different population groups in society. It is therefore imperative to consider not only how best to measure subjective well-being *per se*, but also what other measures should be collected alongside measures of subjective well-being for analytical purposes.

A need for additional information to aid in interpreting and analysing results is not unique to subjective well-being. Most statistical measures are collected alongside, at the least, basic demographic data. Demographics matter to subjective well-being measures just as much as they do to labour market statistics. There are pronounced differences in average levels of subjective well-being across a range of different demographic groups, including based on gender, age and migration status (Dolan, Peasgood and White, 2008). For example, one of the best-known features of life satisfaction data is the "U-shaped" relationship between age and average life satisfaction (Blanchflower and Oswald, 2008). Similarly, there are differences between men and women in life satisfaction and affect measures that are not fully accounted for, even when controlling for income and education (Boarini, Comola et al., 2012).

Beyond demographics, subjective well-being affects and is affected by a wide range of different factors. Material conditions (e.g. income, consumption, wealth) affect subjective well-being (Dolan, Peasgood and White, 2008), but so do factors relating to quality of life. Health status, unemployment, social contact and safety all impact on life satisfaction in important ways (Boarini, Comola et al., 2012). In the context of affect data collected through time-use diaries, it is possible to collect information on an additional range of variables, such as the activity associated with a particular affective state.

Finally, there is a strong case for collecting some additional psychological variables alongside measures of subjective well-being. These include measures of personality type, expectations about the future and views about past experiences. Such measures may help to disentangle fixed effects at the personal level when it is not possible to collect panel data.

The precise range of co-variates to collect alongside measures of subjective well-being will vary with the specific aspect of subjective well-being that is of interest and with the research question being examined. Despite this, it is possible to present some general guidelines on the most important information that should be collected alongside measures of subjective well-being.

Most of the co-variates described below are regularly collected by national statistical agencies, and international standards for their collection do exist. No attempt is made here to specify the details of how these variables should be collected, and it is assumed that existing standards apply. This is not true for a few measures, such as those related to personality, trust and belonging. In these cases some general guidelines are provided. However, as many of these issues (such as the measurement of personality traits) are complex topics in their own right, this chapter does not provide detailed recommendations.

Demographics

Demographic variables cover the basic concepts used to describe the population being measured and to allow the analysis of how outcomes vary by population sub-group. As such, including a range of demographic measures in any attempt to measure subjective well-being is of utmost importance, in particular the following measures:

- ***Age.*** The age of the respondent, in single years if possible. Age bands, while allowing for some cross-classification, are less desirable both because they allow less flexibility with respect to the groups examined, and because they do not facilitate analysis of age as a continuous variable.
- ***Sex or gender.*** The sex or gender of the respondent.

- ***Marital status. The legal marital status*** of the respondent, including whether the respondent is widowed, divorced or separated and the ***social marital status*** of the respondent, including whether the respondent is living as married even if not legally married.
- ***Family type.*** Family type refers to a classification of the respondent's family unit, including whether they are single or living with a partner and whether children are present.
- ***Children.*** The number and age of children in the respondent's family unit, along with relationship to the respondent.
- ***Household size.*** The number of people living in the respondent's household. Household size is a distinct concept from family size, as more than one family unit can live in a dwelling. Household size is essential to allow an understanding of the impact of household income on subjective well-being.
- ***Geographic information.*** While privacy concerns may prevent the release of detailed geographical information relating to the respondent, estimates can be disaggregated by some broad level geographic regions such as urban and rural, capital city, states/provinces, etc. Geo-coding allows for merging with other datasets also containing geo-codes, such as environmental data.

In addition to the demographic measures identified above, which can be considered essential, a number of additional demographic variables may also be desirable to include. The precise relevance of these may, however, vary depending on national circumstances and the research priorities being considered:

- ***Migration status/country of birth/year of arrival.*** Migration status, such as permanent residence, citizenship, etc., and/or country of birth of the respondent.
- ***Ethnic identification.*** The ethnic identity or identities of the respondent may be of high policy importance in ethnically diverse societies.
- ***Language.*** The primary language of the respondent. It may also be desirable, in some circumstances, to collect information on other languages spoken. Proficiency in the main language of the country in which the survey is taking place may also be important for some purposes.
- ***Urbanisation.*** The classification of the area in which the respondent lives in terms of degree of urbanisation.

Material conditions

The term "material conditions" is used here to cover income, wealth and consumption, as well as other aspects of the material living circumstances of the respondent. Much of the interest in measures of subjective well-being has been focused on the relationship between the material conditions of the respondent and their level of subjective well-being. Traditionally, income has been a major focus. The so-called "Easterlin paradox", described by Richard Easterlin (1974), notes that a rise in household income leads to higher subjective well-being for individuals in the household, but that a rise in average incomes for a country appears not to give rise to a corresponding increase in the country's average subjective well-being. Understanding this apparent paradox is important given the degree to which much of policy is focused on economic growth. There are a number of possible explanations for the paradox, but one is lack of high-quality data linking measures of subjective well-being to the household income of the respondent.

Including income measures in surveys of subjective well-being is essential, and should be considered as important as basic demographic variables. It is important that the income measures used are of high quality, and ideally relate to a relatively long period of time (such as a year). Measures of subjective well-being are likely to be more sensitive to changes in long-term income levels than to short-term fluctuations in weekly or monthly income. The relationship between income (at both aggregate average and individual level) and subjective well-being is log linear (i.e. shows diminishing returns) for measures of life evaluation (Sacks, Stevenson and Wolfers, 2010), and the relationship between income and positive affect in a US sample flattens off entirely (Kahneman and Deaton, 2010). This suggests that, if income is collected in bands rather than as a continuous variable, the bands will need to be narrower in currency terms (but constant in proportionate terms) at lower levels of income than at higher levels:

- ***Income.*** Household income is of greater importance than individual income, since it is household income that drives living standards and consumption possibilities (Stiglitz, Sen and Fitoussi, 2009). However, where it is possible, capturing information on individual and household income is also of interest. In both cases, it is desirable to have information on net (post-tax and transfers) income as well as gross income, and equivalised household income should also be available (to take account of household size and composition).[3] Space permitting, information on the source of the income (wages and salary, capital and investment earnings, government transfers) may also be of interest.
- ***Expenditure and consumption.*** Income flows are a relatively limited measure for the actual level of consumption that a household can support. People may draw on previously accumulated assets or run up debt to smooth consumption over time. Thus, for exploring the relationship between consumption and subjective well-being it is desirable to have measures of expenditure and/or access to specific goods and services. Such measures may perhaps allow for separating living standards (consumption) from status and rank effects (income). Questions on financial stress or the ability to access a given amount of money in an emergency may also be valuable for analytical purposes.
- ***Deprivation.*** Because of the difficulty and costs associated with collecting high-quality information on expenditure, surveys are often under pressure with respect to space available for additional questions. Measures of material deprivation provide an alternative to detailed expenditure data as a way of assessing adequacy of consumption.[4] Because such measures impose a much smaller respondent burden than collecting detailed expenditure data, material deprivation can be a useful way to collect information on consumption so as to inform analysis of the relationship between low material living standards and subjective well-being.
- ***Housing quality.*** Housing quality is an important element of the material conditions in which people live, and there is evidence that housing conditions affect subjective well-being (Oswald et al., 2003). Where the impact of material conditions on subjective well-being is a significant part of the research question, collecting information on housing quality will be important. Key dimensions of housing quality to be collected might include number of rooms, housing costs and specific aspects of quality such as dampness or noise.[5] Data on the number of rooms can be used alongside household composition information to assess overcrowding, while housing costs may be used to measure income net of housing costs.

Measures of subjective well-being and household economic statistics are complements rather than substitutes. Understanding the relationship between subjective well-being and economic variables such as income and expenditure is an important rationale for collecting subjective measures in the first place. Collecting data on the economic and social determinants of subjective well-being can then permit their relative importance to be established. Careful thought should be given to facilitating this analysis, not only by including measures of income and wealth in surveys focused on well-being, but also – space and cost permitting – by placing measures of subjective well-being in household income and expenditure surveys and in employment-related surveys.

Quality of life

Quality of life is a broad term covering those aspects of overall well-being that are not captured only by material conditions. The Sen/Stiglitz/Fitoussi commission described quality of life as comprising "the full range of factors that influences what we value in living, reaching beyond its material side" (Stiglitz, Sen, Fitoussi, 2009). Information on these factors is important when measuring subjective well-being, because they are strongly correlated with subjective well-being even after controlling for income and demographic factors (Helliwell, 2008; Dolan, Peasgood and White, 2008; Boarini et al., 2012). In fact, it is likely that much of the simple correlation between individual income and subjective well-being occurs only because income is itself correlated with some measures of quality of life. Evidence of this can be found in the fact that the size of the coefficient on income decreases sharply when quality-of-life measures are included in a regression model (Boarini et al., 2012). This suggests either that higher income is one of the channels through which the quality of life has been improved or that there are other factors that improve both incomes and quality-of-life measures.[6]

Measurement of some aspects of quality of life is less developed than in the case for income, and it is therefore not possible to point to internationally accepted standards for some areas of quality of life that could be collected alongside measures of subjective well-being. In addition, the range of concepts covered by the notion of "quality of life" is so broad that an attempt to be comprehensive in identifying potential co-variates of subjective well-being would be prohibitively large. Nonetheless, it is possible to identify some of the key concepts for which measures would be desirable:

- ***Employment status*** – employment status is known to have a large influence on subjective well-being, with unemployment in particular associated with a strong negative impact on measures of life satisfaction (Winkelmann and Winkelmann, 1998) and affect (Boarini et al., 2012). There is also good evidence that measures of satisfaction with work predict subsequent labour market behaviour (Clark, Georgellis and Sanfrey, 1998; Card et al., 2010).
- ***Health status*** – both physical and mental health are correlated with measures of subjective well-being (Dolan, Peasgood and White, 2008), and there is evidence that changes in disability status cause changes in life satisfaction (Lucas, 2007). Although health status is complex to measure in household surveys, there is a large pool of well-developed measures available, such as the health state descriptions from the World Health Survey (WHO, 2012), or more specialised question modules, such as the GHQ-12 for mental health (Goldberg et al., 1978). Since 2004, a joint work programme of the UNECE, WHO and Eurostat has been engaged in developing common core measures of health status for inclusion in surveys (the Budapest Initiative). When these measures are finalised and become commonly available, they will form a suitable basis for monitoring health outcomes in general population surveys (UNECE, 2009).

- ***Work/life balance*** – there is significant evidence that aspects of work/life balance impact on subjective well-being, in particular commuting (Frey and Stutzer, 2008; Kahneman and Kruger, 2006), and time spent caring for others (Kahneman and Krueger, 2006). Relevant measures include hours worked (paid and unpaid), leisure time, perceived time crunch as well as information on how time is used.
- ***Education and skills*** – education and skills have obvious interest both as variables for cross-classification and because there is good evidence that education is associated with subjective well-being at a bivariate level (Blanchflower and Oswald, 2011; Helliwell, 2008). In analyses that control for additional factors, such as income and social trust, the correlation falls, suggesting that education may affect subjective well-being partly through its impact on other intermediate variables. The highest qualification attained and years of schooling may be used to measure education and skills. There may also be some value in collecting information on current engagement with education.
- ***Social connections*** – social contact is one of the most important drivers of subjective well-being, as it has a large impact both on life evaluations and on affect (Helliwell and Wang, 2011b; Kahneman and Krueger, 2006; Boarini et al., 2012). Although only some elements can be measured well in the context of general household surveys, measures of human contact, such as frequency of contact with friends and family, volunteering activity, and experience of loneliness, should also be collected where possible.
- ***Civic engagement and governance*** – generalised trust in others as well as more domain-specific measures of neighbourhood and workplace trust are crucial factors when accounting for variation in subjective well-being (Helliwell and Wang, 2011b) and should be collected. More generally, corruption and democratic participation have been shown to affect life evaluations (Frey and Stutzer, 2000), and measures of these concepts are of interest.
- ***Environmental quality*** – environmental quality is inherently a geographic phenomenon, and integrating datasets on environmental quality with household level data on life satisfaction is costly. Nonetheless, there is some evidence that noise pollution (Weinhold, 2008) and air pollution (Dolan, Peasgood and White, 2008) have a significant negative impact on life satisfaction. Silva, De Keulenaer and Johnstone (2012) also show that subjective satisfaction with air pollution is correlated with actual air pollution. To understand the impact of environmental quality on subjective well-being, it will be important to link actual environmental conditions to reported subjective well-being via geo-coding. Particular issues of concern include air quality and the extent of local green space.
- ***Personal security*** – security is important to subjective well-being. This is reflected in correlations between experience of victimisation and subjective well-being at the individual level (Boarini et al., 2012), as well as by subjective perceptions of safety. For example, living in an unsafe or deprived area is associated with a lower level of life satisfaction, after controlling for one's own income (Dolan, Peasgood and White, 2008; Balestra and Sultan, 2012). Measures of experience of victimisation and perceived safety should both be collected, as is already done in standard victimisation surveys, because subjective well-being appears to be more strongly affected by perceived crime rates than by actual rates (Helliwell and Wang, 2011b).

Psychological measures

Personality type has a significant impact on how people respond to questions on subjective well-being (Diener, Oishi and Lucas, 2003; Gutiérez et al., 2005). While this will not normally bias results if personality is uncorrelated with the main variables used in the analysis of subjective well-being, it is desirable to control for it if possible. In panel surveys, personality type can be controlled for, to some extent, using individual fixed effects. In cross-sectional household surveys this is not possible. One approach is to incorporate measures of personality type such as the standard instrument for the Five Factor Model (Costa and McCrae, 1992) in surveys focusing on subjective well-being.[7] Although such measures are rarely used in official statistics, this is an area that may warrant further investigation.

Aspirations and expectations, which form part of the *frame of reference*[8] that individuals use when evaluating their lives or reporting their feelings, are also of interest when analysing data on subjective well-being. There is good evidence that life evaluations are affected by aspirations (Kahneman, in Kahneman, Diener and Schwarz, 1999), and it has been suggested that differing aspirations may account for some cultural differences in life evaluations (Diener, Oishi and Lucas, 2003). There is less evidence with respect to how aspirations impact on measures of affect or eudaimonia. Nonetheless, information on people's aspirations and expectations would be useful for investigating this relationship. There are no standard approaches to measuring aspirations and expectations, so it is not possible to be specific as to best practice in approaching measures of this sort. However, this area is one where further research would be of high value.

Time-use diaries

Although all of the measures identified as relevant to household surveys remain equally relevant to time-use surveys, the use of time diaries opens the way to collect additional co-variates not possible in standard household surveys. This is particularly the case where information on aspects of subjective well-being, such as affect, is collected in the diary itself. Several implications of collecting subjective well-being measures via time-use diaries are worth noting specifically:

- ***Activity classification*** – the standard activity classifications (Eurostat, 2004) are central to time-use diaries, and are of primary importance in interpreting information on subjective well-being.
- ***With whom*** – there is evidence that whether an activity is performed alone or with others, and the respondent's relationship to the others, are important to subjective well-being (Kahneman and Krueger, 2006). This reinforces the value of collecting information on "with whom" an activity took place where subjective well-being measures are collected.
- ***Location*** – the location of the activity in question and the impact that this has on subjective well-being is little researched. However, such information potentially brings useful context to analysis, and should be collected alongside subjective well-being and activity classification where possible.
- ***For whom*** – permits disaggregation of activity data by the *purpose* of the activity. This allows for analysing activities done for voluntary organisations, persons with a disability, family and non-family members, which may all have useful analytical possibilities for subjective well-being.

2. Survey and sample design

One important distinction between measures of subjective well-being and many of the measures typically included in official statistics is that subjective well-being measures will almost invariably need to be collected through sample surveys. In contrast to many economic or population statistics, there is generally no administrative database that would produce subjective information without, in effect, incorporating survey questions in an administrative process.[9] Thus, issues relating to survey and sample design are fundamental to producing trustworthy and reliable measures of subjective well-being.

It is not the role of this chapter to provide detailed guidelines on sample frames and sample design. These are specialist areas in their own right, and excellent guides exist for data producers who are seeking advice on these technical aspects of data collection (UN, 1986). However, in survey design, as in other aspects of design, form should follow function. The fact that subjective well-being is the goal of measurement has implications for survey design. This section discusses some of the most significant considerations for the measurement of subjective well-being with respect to the target population, to when and how frequently the data should be collected, to what collection mode should be used, and to identifying the most appropriate survey vehicle.

Target population

The target population for a survey describes the complete set of units to be studied. A sample survey will generally attempt to achieve a representative sample of the target population. However, the target population may be more detailed than the total population from which the sample is drawn. It may also specify sub-populations that the survey describes. For example, the total population might be all persons aged 15 and over living in private dwellings in a specified area. However, the target population might also specify males and females as sub-populations of interest, requiring the sampling frame to accommodate distinct analysis of these two groups. More generally, sub-groups are often defined by such characteristics as age, gender, ethnicity, employment status or migrant status.

Some surveys with the household as the unit of measure rely on a single respondent (such as the head of household) to provide responses for the household as a whole. This cannot be used for measures of subjective well-being, since the cognitive process of evaluating and responding with respect to one's own subjective well-being is very different to that of providing an estimate of another householder's state of mind. Responses to questions on subjective well-being are inherently personal, and consequently the unit of measure for subjective well-being must be the individual. This implies that the sampling frame must produce a representative sample of individuals or households as if all individuals are personally interviewed. While this will typically not be an issue for surveys where the individual is the primary unit of analysis, some household surveys may require an additional set of individual weights to derive individual estimates. Surveys where the response is on the basis of "any responsible adult" will in particular be problematic in this regard.

The target age group for measures of subjective well-being will vary with respect to the goals of the research programme. For example, in the context of research on retirement income policies, it may be appropriate to limit the target population to persons aged 65 or older. In general, however, measures of subjective well-being would usually be collected for all the adult population (aged 15 years and older).

Children

Child well-being is a significant policy issue, both because child well-being has an important impact on later adult outcomes (OECD, 2009), and because the well-being of children is important in its own right. In the analysis of child outcomes, parental or household responses are often used as proxies for the situation of children. In many cases this is a reasonable assumption. Household income, for example, is obviously much more relevant to the living circumstances of a child than would be the income earned by the child themselves. However, this is not true for measures of subjective well-being. If there is a policy interest in the subjective well-being of children, providing data on this will necessarily involve obtaining responses from children. While parents may be able to provide a second-person estimate of the well-being of their child, this is a conceptually different construct from the child's own subjective well-being.

National statistical agencies rarely collect information from respondents younger than 15 or 18 years-old. This reflects both legal issues and concern about the acceptability of such practices to respondents. These are real and significant issues, and must be treated accordingly. However, it is important to note that the available evidence suggests that children are capable of responding effectively to subjective well-being questions from as young as age 11 with respect to measures of life evaluation and affective state (UNICEF, 2007).

While there may be ethical issues with interviewing young children, it is also important to consider the implications of not including the voices of children when measuring subjective well-being. As the focus of this chapter is on general population surveys, questions focused specifically at young children are therefore not provided. However, this remains a significant gap that future work should address.

People not living in private households

One population group that may be of high policy interest, but which is not typically covered in household surveys, is people not living in private households. This group includes people living in institutions, including prisons, hospitals or residential care facilities, as well as people with no fixed residence, such as the homeless. These groups raise two issues with respect to the measurement of subjective well-being. The first problem is common to all attempts to collect statistical information on such groups – that such population groups tend to be excluded from standard household survey sample frames. This means that, at a minimum, specific data collection efforts will be required based on a sample frame designed to cover the relevant institutions. In some cases, such as for the homeless, it may be difficult to develop any statistically representative sampling approach at all.

A more significant challenge faced in the measurement of subjective well-being is that many of the people in the relevant groups may not be able to respond on their own behalf. This is particularly the case for people institutionalised for health-related reasons that affect mental functioning (including people with some mental illnesses, or with physical illnesses limiting the ability to communicate, and the very old). In these cases it is not possible to collect information on a person's subjective well-being. Proxy responses, which might be appropriate for some types of data (income, marital status, age), are not valid with respect to subjective well-being.

Frequency and duration of enumeration

The frequency with which data is collected typically involves a trade-off between survey goals and available resources. All other things being equal, more frequent collection of data will improve the timeliness of estimates available to analysts and policy-makers, and will make it easier to discern trends in the data over time. More frequent enumeration, however, is more costly both in terms of the resources involved in conducting the data collection and in terms of the burden placed upon respondents. It is therefore important that decisions around the frequency of data collection are made with a clear view to the relationship between the timeliness and frequency of the data produced and the goals of the data collection exercise.

It is not possible to provide specific guidelines for how frequently measures of subjective well-being should be collected covering every contingency, since the range of possible data uses is large and the frequency at which data are needed will vary depending on the intended use and on the type of measure in question. However, some general advice can be provided. Aggregate measures of subjective well-being generally tend to change only slowly over time. This reflects the relatively slow movements in most of the social outcomes that affect subjective well-being and the fact that many changes only impact on a small proportion of the population. For example, unemployment – which is associated with a change of between 0.7 and 1 on a 0 to 10 scale (Winkelmann and Winkelmann, 1998; Lucas et al., 2004) – typically affects between three and 10% of the adult population. Thus, even a large shift in the unemployment rate – say, an increase of 5 percentage points – will translate only into a small change in measures of subjective well-being (Deaton, 2011).

The relatively slow rate of change in measures of subjective well-being might appear to suggest that such measures do not need to be collected frequently. However, the small absolute size of changes in subjective well-being also means that standard errors tend to be large relative to observed changes. A number of observations are therefore needed to distinguish between a trend over time and noise in the data. Box 3.1 illustrates this point. For this reason, despite (or indeed, because of) the relatively slow rate of change in subjective well-being data, it is desirable that measures are collected on a regular and timely basis. For the most important measure used in monitoring well-being, an annual time series should be regarded as the essential minimum in terms of frequency of enumeration. More frequent monthly or weekly data is, however, likely to be of lower value (Deaton, 2011). (It should be pointed out that frequent, or rolling sample, surveys increase the possibilities for identifying the causal impacts of other factors whose dates can be identified. It was only the daily frequency of observations that made it so easy to discover and eliminate the question-order effects in Deaton (2011).

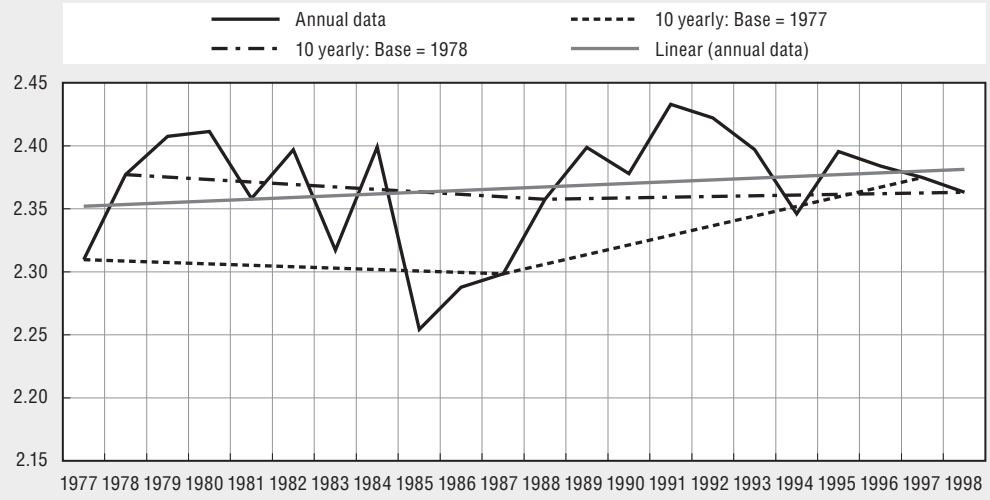
Duration of enumeration

The duration of the enumeration period (i.e. the period of time over which information is collected) is immensely important for measures of subjective well-being. Unlike measures of educational attainment or marital status, for which it does not usually matter at what point during the year the data are collected, the precise timing of the collection period can have a significant impact on measured subjective well-being (Deaton, 2011). For example, measures of positive affect are higher on weekends and holidays than on week days (Helliwell and Wang, 2011a; Deaton, 2011).

Box 3.1. **Identifying trends in time series of subjective well-being: Implications for frequency of measurement**

It might seem logical that, if measures of subjective well-being change only slowly over time, they will need to be measured only infrequently. Figure 3.2 illustrates why this assumption does not necessarily hold, by plotting the time series changes of life satisfaction in the Netherlands using data collected in the Eurobarometer. Because the Eurobarometer has a relatively small sample size, the error is large relative to the size of changes over time. Between 1977 and 1997, life satisfaction in the Netherlands was largely static. Despite large fluctuations in the annual time series, the trend line for this 20-year period is almost flat.

Figure 3.2. **Life satisfaction in the Netherlands 1977-97: Eurobarometer**



Source: Eurobarometer.

Two additional lines included in Figure 3.1 illustrate the picture that would emerge if information had been collected only every ten years. If the base year were 1977, then a 10-yearly data collection would show a slight decline in life satisfaction between 1977 and 1987, followed by a substantial increase from 1987 to 1997. The net effect over twenty years is seen to be a significant increase in life satisfaction. Changing the base year to 1978 produces a different pattern, with life satisfaction declining from 1978 to 1988, before flattening out from 1988 to 1998. The overall effect over 20 years this time would be a slight decrease in life satisfaction. None of these 10-yearly patterns reproduces the pattern highlighted by annual data.

The fact of being sensitive to the point in time at which they are collected is not unique to measures of subjective well-being. Many core labour market statistics, for example, have a pronounced seasonality, and published statistics usually adjust for this. However, such adjustments require collecting data over the course of a whole year in order to produce the information required for seasonal adjustments.

The fact that subjective well-being does vary over the year suggests strongly that a long enumeration period is desirable. Ideally, enumeration would take place over a full year, and would include all days of the week, including holidays. This would ensure that

measures of subjective well-being provide an accurate picture of subjective well-being across the whole year. Where a year-long enumeration period is not possible, enumeration should be spread proportionately over all the days of the week as far as is possible. All days of the week need to be covered because day of the week can impact on the level of subjective well-being reported (Helliwell and Wang, 2011a). Any attempt to measure the "typical" level of subjective well-being for a group would need to account for regular variations over time, and it may be necessary to develop a specific set of weights to ensure that responses from all days contribute equally to the final estimate.

Holidays (and to some degree the incidence of annual leave) are more problematic in that they tend to be distributed unevenly over the course of the year. Thus, if enumeration cannot be spread over a whole year, there is a risk that an incidence of holidays during the enumeration period that is greater or lesser than normal might bias the survey results. For this reason, it is essential in surveys collected with relatively short enumeration periods that the impact of the inclusion of data collected during any holidays is checked. While it may not be necessary to omit data collected during holidays from output if the impact is negligible or weak, the available evidence on the magnitude of some holidays suggests that testing for potential bias from this source is important. What constitutes a holiday will need to be considered with respect to the context in which the survey is collected. However, it is worth noting that Deaton (2011) finds a large effect for Valentine's Day in the United States, despite the day not being a public or bank holiday.

Sample size

Large samples are highly desirable in any survey, as they reduce the standard error of estimates and allow both a more precise estimate of subjective well-being as well as a greater degree of freedom with respect to producing cross-tabulations and analysis of results for population sub-groups. With measures of subjective well-being, sample size is particularly important because of the relatively small changes in subjective well-being associated with many areas of analytical interest. Deaton (2011), for example, notes that the expected decline in life satisfaction due to the changes in household incomes and unemployment associated with the 2008 financial crisis is less than the standard error on a sample of 1 000 respondents, and only three times larger than the standard error on a sample of 30 000 respondents. Thus, large samples are highly desirable for measures of subjective well-being.

Although it is impossible to give precise guidelines for what is an appropriate sample size, some general criteria can be noted. Most of the factors that should be taken into account in the planning of any survey also apply when collecting information on subjective well-being. Available resources, respondent burden, sample design (a stratified sample will have a different sample size to a random sample with the same objectives, all other things equal), anticipated response rate and the required output will all influence the desirable sample size. The need for sub-national estimates, in particular, will play an important role in determining the minimum required sample.

Over and above this, some features specific to measures of subjective well-being will influence the desired sample size. On a 0-10 scale, an effect size of 1 implies a very large effect when analysing the determinants of subjective well-being (Boarini et al., 2012). Changes over time are even smaller. The analysis of subjective well-being data therefore requires a relatively large sample size in order to achieve the statistical precision required.[10]

Mode

Surveys can be carried out in a number of different modes. Because the mode of collection influences survey costs and respondent burden and can induce mode effects in responses, the choice of mode is an important decision when collecting data. The two modes most commonly used to collect information on subjective well-being are Computer-Assisted Telephone Interviewing (CATI), conducted by an interview over the telephone, and Computer-Assisted Personal Interviewing (CAPI), where the interviewer is personally present when recording the data. Computer-Assisted Self-Interview (CASI) surveys can occur in the presence of an interviewer, when the interviewer is on hand but the respondent enters their own data into a computer questionnaire, or without an interviewer present, such as when the respondent completes an Internet survey. For some purposes traditional chapter-based self-complete surveys are still likely to be relevant. Most time-use diaries, for example, are self-completed chapter diaries filled in by the respondent.

As outlined in Chapter 2, there is good evidence that the collection mode has a significant impact on responses to subjective well-being questions. In general, the use of CASI as a mode tends to produce lower positive self-reports than the use of CAPI, and this is assumed to be because interviewer-led approaches are more likely to prompt more socially desirable responding. CATI is viewed as the least reliable way to collect consistent subjective well-being data, because in these conditions the interviewer is unaware of whether the respondent is answering in a private setting or not, and it is more challenging for interviewers to build rapport with respondents.

As with other features of survey design, the choice of the survey mode will be influenced by a variety of factors, including resource constraints. However, the balance of evidence suggests that, where resources permit, CAPI is likely to produce the highest data quality. This is probably due in part to the rapport that interviewers can build in face-to-face situations. However, CAPI also provides the opportunity to use show cards, which CATI lacks. Show cards that include verbal labels for the scale end-points are particularly valuable in collecting information on subjective well-being where the meaning of the scale end-points changes between questions, as this can impose a significant cognitive burden on respondents (ONS, 2012).

In terms of data quality, CAPI with show cards should be considered best practice for collecting subjective well-being data. Where other modes are used it is important that data producers collect information to enable the impact of mode effects to be estimated. National statistical agencies, in particular, should consider experimentally testing the impact of the mode on responses to the core measures of subjective well-being and publishing the results along with any results from CATI or CASI[11] surveys.

Survey vehicles

Questions on subjective well-being should not typically be the subject of a specific survey. As discussed earlier in this chapter, analytical interest in measures of subjective well-being is commonly focused on the interaction between measures of subjective well-being and measures of objective outcomes, including income, aspects of quality of life and time use. It should also be considered that, in most cases, subjective well-being measures are relatively simple and easy to collect. For example, the UK Office for National Statistics found that the four subjective well-being questions used in the Integrated Household Survey take approximately 30 seconds to complete (ONS, 2011). Even a

relatively comprehensive approach to measuring subjective well-being is likely to be more on the scale of a module that could be added to existing surveys rather than requiring a whole survey questionnaire in itself. A key question to consider then is which survey vehicles are most appropriate to the task of measuring subjective well-being.

It is impossible to provide definitive guidance on this issue, because the range of household surveys collected – even among national statistical agencies – varies significantly from country to country. However, it is possible to identify the roles that different survey vehicles can play in collecting subjective well-being data. Seven classes of survey vehicle are relevant to subjective well-being and meet slightly different needs. These are:

- Integrated household surveys.
- General social surveys.
- Time-use surveys.
- Victimisation surveys.
- Health surveys.
- Special topic surveys.
- Panel surveys.

Integrated household surveys

Integrated household surveys include the primary surveys used by national statistical agencies to collect information on issues such as income, expenditure and labour market status. In some countries information such as this is collected through separate surveys, such as a labour force survey, while other countries, such as the United Kingdom, rely on an integrated household survey with sub-samples focused on particular topics. Another similar example is the EU-SILC, which consists of a core survey focused on income and living conditions alongside a range of special topic modules. The 2013 EU-SILC module is focused explicitly on well-being. Such surveys are generally not appropriate to be the sole source of information on subjective well-being, as they have a clearly-defined focus that may not align well with an extensive module of subjective well-being and space in these surveys is at a premium. However, such surveys may be more appropriate as a vehicle for a limited set of core questions or a primary measure of subjective well-being intended for monitoring purposes. These questions take up relatively little space in a survey and demand both large sample sizes and regular collection in order to support the effective monitoring of outcomes. Further, subjective measures of this sort complement the economic focus of many integrated household surveys by capturing information on the impact of non-economic factors in a relatively compact form.

General social surveys

Not all national statistical agencies run general social surveys, and among those that do, the content and focus vary considerably. Some national statistical agencies, such as the Australian Bureau of Statistics, focus their general social survey primarily on measures of social capital and social inclusion, while others rotate modules on different topics between survey waves (Statistics Canada) or are explicitly multi-dimensional (Statistics New Zealand). The latter two approaches are particularly appropriate vehicles for collecting information on subjective well-being (and indeed, both Statistics Canada and Statistics

New Zealand collect information on subjective well-being in their general social surveys). Surveys with rotating content, such as the Canadian General Social Survey, offer the opportunity for a subjective well-being module that can collect information in some depth if this is determined to be a priority. Surveys with a wider focus, such as the New Zealand General Social Survey, are particularly valuable in that they allow for the analysis of the joint distribution of subjective well-being and of a wide variety of other topics, including material conditions and objective aspects of quality of life. Regardless of whether a specific subjective well-being module is collected as part of a general social survey, it is very desirable that at least the core module be collected in all general social surveys.

Time-use surveys

Time-use surveys typically involve respondents completing a time-use diary alongside a questionnaire on demographic and other information. The inclusion of a time-use diary offers a unique opportunity to gather information that relates activities to particular subjective states and to collect information on the amount of time spent in different subjective states. In particular, time-use surveys have been used to collect data on affect at varying levels of detail. The American Time Use Survey 2011 included an implementation of the Day Reconstruction Method (Kahneman and Krueger, 2006), which collected detailed information on the affective states associated with a representative sample of episodes drawn from the diaries. This allows analysis of how different affective states vary depending on activity type and calculation of the aggregate amount of time spent in different affective states. The *Enquête Emploi du temps 2010*, run by the French statistical agency, the INSEE, uses an alternative approach to collecting information on subjective well-being in time diaries. Rather than collecting detailed information on a sample of episodes, the INSEE selected a sub-sample of respondents to self-complete a simple seven point scale (-3 to +3), rating each activity from *très désagréable* to *très agréable*. This gives less information on each activity for which information is collected but gathers information on all recorded diary time, therefore providing a larger effective sample of diary entries with subjective well-being information attached.

Victimisation surveys

Victimisaton surveys collect information on the level and distribution of criminal victimisation in a society. They are intended to answer questions such as how much crime takes place, what are its characteristics, who are its victims, whether the level of crime is changing over time, who is at risk of becoming a victim, and how do perceptions of safety relate to the actual risk of victimisation (UNECE, 2010). The interaction between victimisation, perceptions of safety and subjective well-being is of high interest, both from the perspective of understanding how victimisation affects well-being, and in order to better understand the impact on the victim of different types of victimisation. Subjective well-being questions are thus of high value to such surveys.

Health surveys

Health surveys already have a considerable tradition of the inclusion of measures of subjective well-being as part of overall and mental health modules such as the widely used GHQ-12 and SF-36 modules. These include questions relating to all three aspects of subjective well-being. However, these modules are calibrated for a specific purpose – measuring overall health status or pre-screening for mental health issues – and in many ways do not conform

with best practice in measuring subjective well-being as outlined here. Because of the importance of health status to subjective well-being, there is considerable value in adding a small number of specific subjective well-being measures to such surveys where possible.

Special topic surveys

Many national statistical agencies run one-off or periodic special topic surveys that are intended to explore a topic in greater detail than would be possible through a question module in a regular survey. Because the content of such a survey can be tailored to the topic in question, such surveys are excellent vehicles for exploring aspects of subjective well-being in more depth. Issues relating to the relationship between different aspects of subjective well-being (i.e. life evaluation, affect, eudaimonic well-being), and between single-item and multiple-item measures of subjective well-being can be examined with such data. However, because of the "one-off" nature of such surveys (or the long periodicity associated with such surveys when they do repeat), special topic surveys are less appropriate for monitoring well-being over time.

Panel surveys

Panel surveys follow the same individuals over time, re-interviewing them in each wave of the survey. Because of this, panel surveys are able to examine questions of causality in a way that is not possible with cross-sectional surveys. Both the German Socio-Economic Panel (GSOEP) and Understanding Society (formerly the British Household Panel Survey) have included questions on subjective well-being for some time, and much of the evidence on the nature of the relationship between life evaluations and their determinants derives from these surveys.

3. Questionnaire design

Questionnaire design is an iterative process involving questionnaire designers, those responsible for determining survey content, and data users. A questionnaire designer must balance the cognitive burden on the respondent, a limited time budget for the survey, and the need to have a questionnaire that is clear, comprehensible and flows well, with different (and often competing) data needs. It is neither possible nor desirable for this chapter to provide a single questionnaire on subjective well-being for users to implement. Instead, the intent of this section is to provide a set of tools to support the development of surveys containing questions on subjective well-being rather than to prescribe a single approach to its measurement.

Some general guidance on issues affecting the inclusion of measures of subjective well-being into a survey is provided below. In particular, the issues of question placement and translation are discussed on their own. This is accompanied by a set of prototype question modules that questionnaire designers should adapt to the specific conditions under which they are working. This section also describes the rationale behind the question modules and an explanation of the template used to describe them. The question modules are attached to these guidelines as Annex B (A to F).

Question placement

Question order and the context in which a question is asked can have a significant impact on responses to subjective questions (see Chapter 2). Although measures of subjective well-being are not uniquely susceptible to such effects – question order and

context will impact on all survey responses to some extent – the effect is relatively large in the case of subjective well-being. Several well-known examples suggest that such effects do need to be taken into account when incorporating questions on subjective well-being into a survey.

In general, question order effects appear to occur, not because the question was early or late in the questionnaire *per se*, but because of the contextual impact of the immediately preceding questions. Thus, the key issue is to identify the most effective way to isolate questions on subjective well-being from the contextual impact of preceding questions. The most direct way of managing contextual effects of this sort is to put subjective questions as early in the survey as possible. Ideally, such questions should come immediately after the screening questions and household demographics that establish respondent eligibility to participate in the survey. This practice almost eliminates the impact of contextual effects and ensures that those that cannot be eliminated in this way are consistent from survey to survey.

However, this cannot be a general response to the issue of dealing with contextual effects for two reasons. First, there will be instances when questions on subjective well-being are added to well-established surveys. In these conditions, changing the flow of the questionnaire would impose significant costs in both resources and data quality. Introducing questions on subjective well-being early in such a survey might ensure that contextual effects do not impact the subjective questions, but this would come at the expense of creating significant contextual effects for the following questions. Second, in cases where there are several such questions in the survey, they cannot all be first.

With these factors in mind, four key recommendations emerge with regard to the placement of subjective well-being questions in surveys. These are as follows:

- *Place important subjective well-being questions near the start of the survey.* Although, as noted above, placing questions early in a survey does not eliminate all of the problems associated with context effects, it is the best strategy available and should be pursued where possible. In particular, for the core measures of subjective well-being, for which international or time series comparisons are an important consideration, it is desirable to place the questions directly after the initial screening questions that result in a respondent's inclusion in the survey. The core measures module included as an annex to this chapter is intended to be placed at the start of a survey in this way.
- *Avoid placing the subjective well-being questions immediately after questions likely to elicit a strong emotional response or that respondents might use as a heuristic for determining their response to the subjective well-being question.* This would include questions on income, social contact, labour force status, victimisation, political beliefs or any questions suggesting social ranking. The best questions to precede subjective questions might be relatively neutral factual demographic questions.
- *Make use of transition questions to refocus respondent attention.* One technique that has been used to address contextual effects resulting from a preceding question on a subjective well-being question is using a transition question designed to focus the respondent's attention on their personal life. Deaton (2011) reports that the introduction of such a question in the Gallup Healthways Well-being Index in 2009 eliminated over 80% of the impact from a preceding question on politics on the subsequent life evaluation measure.[12] However, it is important to consider the risk that transition questions might introduce their own context effects. For example, drawing attention to a respondent's

personal life may lead them to focus on personal relationships or family when answering subsequent questions about life overall. Development of effective transition questions should be a priority for future work.

- *Use of introductory text to distinguish between question topics.* Well-worded text that precedes each question or topic can serve as a buffer between measures of subjective well-being and sensitive questions. However, there is little hard evidence on the degree of effectiveness or optimal phrasing of such introductory text. A standard introductory text has been included in each of the prototype question modules included as an annex to this chapter. This text is based on what is believed to be best practice. Consistent use of it should help reduce context effects (and will eliminate bias caused by inconsistent introductory text). Further cognitive testing or experimental analysis of the impact of different types of introductory text would, however, be of high value.

Question order within and between subjective well-being modules

Questions on subjective well-being can be affected by previous subjective well-being questions just as easily as by questions on other topics. This has implications for the structure of subjective well-being question modules (particularly where more than one aspect of subjective well-being is addressed), as well as for the presentation of questions within modules and whether it is advisable to include several questions that address very similar topics (see Chapter 2).

In terms of ordering question modules themselves, overall the evidence suggests that moving from the general to the specific may be the best approach. This implies that overall life evaluations should be assessed first, followed by eudaimonic well-being, with more specific questions about recent affective experiences asked next and domain-specific questions last. This is because domain-specific measures in particular risk focusing respondent attention on those domains included in the questions, rather than thinking about their lives and experiences more broadly.

Question order within a battery of questions can also be important – particularly where a group of questions include both positive and negative constructs (such as in the case of affect and some measures of eudaimonia). Although full randomisation of such questions may be optimal, in practice switching between positive and negative items may prove confusing for respondents, who may deal more easily with clusters of questions of the same valence. As discussed in Chapter 2, more evidence is needed to resolve this trade-off, but in the meantime, consistency in the presentation approach (whether randomised or clustered) across all surveys will be important, particularly in terms of whether positive or negative constructs are measured first. In the question modules attached to these guidelines, a clustered approach has been adopted.

Finally, asking two questions about a very similar construct can be confusing for respondents, leading them to provide different answers because they anticipate different answers must be required of them. This means that including several very similar questions about life evaluations, for example, could mean respondents react differently to these questions than when each question is presented in isolation. Thus it is important to have consistency in the number of measures used to assess a given construct, and the order in which those measures are used.

Translation

The exact question wording used in collecting information on subjective well-being matters a lot for responses. As discussed in Chapter 2, a standardised approach to question wording is important for comparisons over time or between groups. This is relatively straight-forward where all surveys are in a single language. However, international comparisons or studies in multi-lingual countries raise the issue of translation. This is a non-trivial matter. Translating survey questionnaires to work in different languages is challenging for any survey, and the potential sensitivity of subjective well-being questions to differences in wording only reinforces this issue.

Potential issues arising from translation cannot be entirely eliminated, but they can be managed through an effective translation process. An example of good practice in the translation of survey questionnaires is provided by the *Guidelines for the development and criteria for the adoption of Health Survey Instruments* (Eurostat, 2005). Although focused on health survey instruments, the framework for translation presented there has broader applicability, and is highly relevant to the measurement of subjective well-being. The health survey guidelines identify four main steps to the translation procedure:

- *Initial or forward translation* of the questionnaire from the source document to the target language.
- *Independent review* of the translated survey instrument.
- *Adjudication* of the translated survey instrument by a committee to produce a final version of the translated survey instrument.
- *Back translation* of the final version of the translated survey instrument into the source language.

Most of the best-practice recommendations identified by Eurostat for health surveys also apply with respect to the measurement of subjective well-being. It is desirable that the initial translation be carried out by at least two independent translators who have the destination language as their mother tongue and who are fluent in the source language. Translators should be informed about the goal of the study and be familiar with the background, origin and technical details of the source questionnaire as well as with the nature of the target population. The reviewer at stage 2 should be independent from the translators, but will ideally need a very similar skill set. Both the reviewer and the translators should be on the adjudication panel, along with an adjudicator whose main area of expertise is the study content and objective. As with any survey design, cognitive interviewing and field testing should be undertaken and the results of this reviewed before the full survey goes into the field.

Back translation is somewhat controversial in the literature on survey translation, with some experts recommending it and others not (Eurostat, 2005). The effect of back translation is generally to shift the focus onto literal translation issues rather than the conceptual equivalent of the original instrument. In the case of the measurement of subjective well-being, back translation is strongly advised. This reflects the sensitivity to question wording of subjective well-being measures (see Chapter 2).

Choice of questions

The choice of which questions to use is of critical importance for measuring subjective well-being. Different questions capture different dimensions of subjective well-being and,

as discussed in Chapter 2, the precise question wording can have a non-trivial impact on results. In selecting questions to incorporate into existing survey vehicles, statistical agencies face trade-offs between the time taken to ask any new questions, the potential impact of new questions on responses to existing questions, and the added information gained from the new questions. These trade-offs will come under particularly severe scrutiny if the survey in question refers to an important and well-established concept (e.g. household income or unemployment).

In selecting subjective well-being questions themselves, there is also a trade-off to manage between using existing questions from the literature that will enable reasonable comparability with previous work, and modifying questions or response formats in light of what has been learned about good practice – including the evidence described in Chapter 2. The approach adopted in this chapter is to recommend tried-and-tested questions from the literature first and foremost. Where a variety of approaches have been used in the past, the rationale for selecting between these is explained. Finally, where there is a case for making small alterations to the question wording based on the evidence in Chapter 2, some modifications are proposed.

For statistical agencies already using subjective well-being measures in their surveys, a crucial question will be whether the potential benefit of using improved measures, and/or more internationally comparable measures, outweighs the potential cost of disrupting an established time series. This is a choice for individual statistical agencies, and will depend on a number of factors, including what the current and future intended use of the data set is, how drastic the change may be, and how long the time series has been established for. It is recommended that any changes to existing questions are phased in using parallel samples, so that the impact of the change can be fully documented and examined. This will enable insights into the systematic impact of changes in methodology and provide agencies with a potential method for adjusting previous data sets (e.g. Deaton, 2011).

In recognition of the different user needs and resources available to statistics producers, this chapter does not present a single approach to gathering information on subjective well-being. Instead, six question modules are attached to the guidelines as Annex B (A to F). Each question module focuses on a distinct aspect of subjective well-being. Question Module A contains the core measures for which international comparability is the highest priority. These are measures for which the evidence on their validity and relevance is greatest, the results are best understood, and the policy uses are the most developed. Of all the six question modules, Module A is unique in that it contains both life evaluation and affect measures, and because all national statistical agencies are encouraged to implement it in its entirety. When this is not possible, the primary measure outlined in the module should be used at the minimum. Modules B through to E are focused on specific aspects of subjective well-being. These modules are not intended to be used in their entirety or unaltered, but provide a resource for national statistical agencies that are developing their own questionnaires.

The six modules are listed below, and those which it is recommended that national statistical offices implement are highlighted as *recommended* in order to distinguish them from those modules intended as a *resource* for data producers of all types that are developing more detailed questionnaires.

Recommended:

- A. Core measures.

Resource:

- B. Life evaluation.
- C. Affect.
- D. Eudaimonic well-being.
- E. Domain evaluation.

Recommended for time-use surveys:

- F. Experienced well-being.

A. Core measures

The core measures are intended to be used by data producers as the common reference point for the measurement of subjective well-being. Although limited to a few questions, the core measures provide the foundation for comparisons of the level and distribution of life evaluations and affect between countries, over time and between population groups.

Data producers are encouraged to use the core measures in their entirety. The whole module should take less than 2 minutes to complete in most instances. It includes a basic measure of overall life evaluation and three short affect questions. A single experimental eudaimonic measure is also included.

There are two elements to the core measures module. The first is a primary measure of life evaluation. This represents the absolute minimum required to measure subjective well-being, and it is recommended that all national statistical agencies include this measure in one of their annual household surveys.

The second element consists of a short series of affect questions and an experimental eudaimonic question. The inclusion of these measures complements the primary evaluative measure both because they capture different aspects of subjective well-being (with a different set of drivers) and because the difference in the nature of the measures means that they are affected in different ways by cultural and other sources of measurement error. While it is highly desirable that these questions are collected along with the primary measure as part of the core, these questions should be considered a lower priority than the primary measure. In particular, the inclusion of the eudaimonic measure in the core should be considered experimental.

There are essentially two candidate questions for the primary measure. These are the Self-Anchoring Striving Scale (the Cantril Ladder) and a version of the commonly-used question on satisfaction with life. Both have been widely used and have an extensive literature attesting to their validity and reliability. Both questions focus on the evaluative aspect of subjective well-being and have been used in large-scale surveys across many different nations and cultures. The choice between the two measures comes down to a balancing of the strengths and weaknesses of each measure.

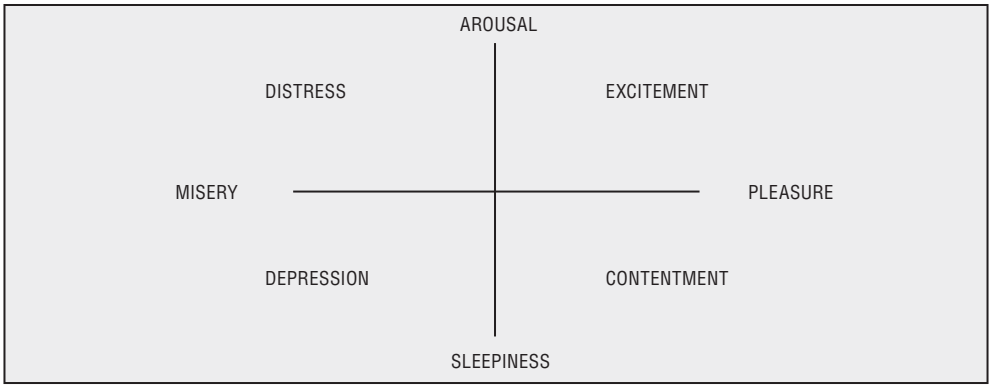
The Cantril Ladder is designed to be "self-anchoring", and is therefore thought to be less vulnerable to interpersonal differences in how people use the measurement scale. In addition, the anchoring element of the scale is explicitly framed relative to the respondent's aspirations. This has led some authors to suggest that it may be more rather than less vulnerable to issues of cross-country comparability (Bjørnskov, 2010). Also, the Cantril Ladder tends to produce a marginally wider distribution of responses than does satisfaction with life. However, the Cantril Ladder is a relatively lengthy question, requiring some explanation of the "ladder" concept involved.

By way of contrast, the satisfaction with life question is simple and relatively intuitive. Compared to the Cantril Ladder, the satisfaction with life question has been the subject of much more analysis, reflecting its inclusion not just in the World Values Survey, but also in crucial panel datasets such as the German Socio-Economic Panel and the British Household Panel Survey.

The Cantril Ladder and the satisfaction with life question are relatively similar in terms of their technical suitability for use as an over-arching measure, particularly if both use the same 11-point (0 to 10) scale.[13] Given this situation, the primary measure included in the core module is a variant of the satisfaction with life question using a 0-to-10 scale. The decisive factor in favour of this choice is the relative simplicity of the question, which will make it easier to incorporate in large-scale household surveys where respondent burden is a significant issue.

Several affect questions are included in the core module. This is because affect is inherently multi-dimensional and no single question can capture overall affect. The various dimensions of affect can be classified in two ways. One of these relates to positive versus negative emotions, while the other relates to level of "arousal". This gives four affect quadrants and is known as the Circumplex model (Larson and Fredrickson, 1999).[14] Figure 3.3 illustrates the Circumplex model. The quadrants are: positive low arousal (e.g. contentment); positive high arousal (e.g. joy); negative low arousal (e.g. sadness); and negative high arousal (e.g. anger, stress). A good measure of affect might attempt to cover all four quadrants.

Figure 3.3. **The circumplex model of affect**



Source: Derived from Russell (1980).

Unlike overall life satisfaction, there is not an obvious choice of a simple affect measure that is suitable for inclusion in general household surveys. Most affect scales have been developed in the context either of the measurement of mental health or of more general psychological research. In the former case, many of the existing scales focus excessively on negative affect, while in the latter the scales may be too long for practical use in a household survey. One model for collecting affect measures in a household survey is provided by the Gallup World Poll, which contains a range of questions on affect covering enjoyment, worry, anger, stress and depression, as well as some physical indicators such as smiling or experiencing pain. These questions now have a significant history of use and analysis behind them (Kahneman and Deaton, 2010). A very similar set of questions (on positive affect only) was proposed by Davern, Cummins and Stokes (2007).

The affect questions contained in the proposed prototype module are based on those in the Gallup World Poll and proposed by Davern, but reduced to a list of two questions covering both the negative quadrants of the Circumplex model of affect and a single positive affect question. Only a single positive question is used because the different aspects of positive affect are, in practice, relatively closely correlated. The moods proposed for measurement are happy, worried and depressed. In each case, a 0-to-10 frequency scale is used for responses (ranging from "not at all", to "all of the time", similar to the scale anchors used in the European Social Survey).

The eudaimonic question is based on a question trialled by the ONS: "to what extent do you feel the things you do in your life are worthwhile?" There is good evidence from the ONS data that this question captures information not covered by life evaluation and affect measures (NEF, 2012). In addition, a similar question was included in the American Time Use Survey well-being module (Krueger and Mueller, 2012). The question proposed here is similar to that used by the ONS. However, because there is as yet no over-arching theory linking individual questions such as the one proposed to "eudaimonia" as a broad concept, the question should be regarded as experimental.

B. Life evaluation

The life evaluation module is not intended to be used in its entirety. To some degree, the measures it contains should be considered as substitutes for each other rather than as complements. Nonetheless, all of the measures included in the module add something over and above the basic satisfaction with life question contained in Module A. Broadly speaking, there are three groups of question contained in Module B.

The first two questions (the Cantril self-anchoring striving scale and the overall happiness question) are alternative measures of the same underlying concept as satisfaction with life. Although there is some debate as to whether the measures do indeed capture exactly the same concept (Helliwell, Layard and Sachs, 2012) or whether the Cantril scale is a more "pure" measure of life evaluation and overall happiness somewhat more influenced by affect (Diener, Kahneman, Tov and Arora, in Diener, Helliwell and Kahneman, 2010), there is no doubt that the measures are all predominantly evaluative. As discussed above, the Cantril scale is somewhat more awkwardly worded than the satisfaction with life question, but tends to produce a slightly wider distribution of responses (ONS, 2011) and has been thought to have a stronger association with income (Helliwell, 2008; Diener, Kahneman, Tov and Arora, in Diener, Helliwell and Kahneman, 2010). However, when the Cantril Ladder and life satisfaction questions are asked of the same respondents, they show essentially identical responses to income and other variables (Chapter 10 of Diener, Helliwell and Kahneman, 2010), so much so that an average of life satisfaction and the Cantril Ladder performs better than either on its own. Some authors have noted that the word "happiness" can be challenging to translate effectively (Bjørnskov, 2010; Veenhoven, 2008), and Bjørnskov further argues that life satisfaction is easier to translate more precisely. However, happiness may be easier to communicate to the public than the more "technical" satisfaction measures. Helliwell, Layard and Sachs (2012) note, based on analysis of the European Social Survey, that averages of overall happiness and life satisfaction perform better than either does alone in terms of the proportion of variance that can be explained by a common set of explanatory variables.

Some of these questions (B3 and B4) capture information on the respondent's perceptions of prior life satisfaction and their anticipated future life satisfaction. This potentially provides some information about how optimistic or pessimistic the respondent feels, but it can also add information on the respondent's overall life evaluation, as a person's expectations of the future are part of how they evaluate their life. This view is reflected in the methodology for life evaluation used by the Gallup Healthways Well-being Index, which is calculated as the average of the Cantril scale and the anticipated Cantril scale 5 years in the future (Gallup, 2012).

Finally, the module includes the five questions (B5 to B9) that together define the Satisfaction With Life Scale (SWLS) developed by Ed Diener and William Pavot. The SWLS is one of the best-tested and most reliable multi-item scales of life evaluation. Since its development in 1985, the SWLS has accumulated a large body of evidence on its performance and has been tested in a number of different languages (Pavot and Diener, 1993). Because the SWLS is a multi-item measure, it has a higher reliability than single-item measures and is more robust to inter-personal differences in scale interpretation than they are. The SWLS adds value to the primary life evaluation measure in contexts where more space is available in a questionnaire, and where a more reliable measure of life evaluation would help interpret and calibrate the results from the primary measure.

C. Affect

Best practice for collecting data on directly-experienced affect involves either sampling people throughout the course of the day and recording their affective state (experience sampling method or ESM) or a detailed reconstruction of daily activity and of the associated affective states (the day reconstruction method or DRM). The former approach (ESM) is not discussed here in detail as it typically involves the use of electronic pagers or similar devices more suited to experimental research design than to official statistics. The DRM approach can be implemented in large-scale surveys containing a time-use diary and forms the basis of the experienced well-being module presented in this chapter. However, time-use diaries are expensive to collect and code, and there are times when it may be desirable to collect affect data from a general household survey. This module provides an approach to collecting affect data in such a survey and expands on the more limited range of affect questions contained in the core questions module.

There are several approaches to measuring affect in household surveys. The European Quality of Life Survey (Eurofound, 2007), for example, asks five questions about how people felt during the previous two weeks. These questions ask respondents to rate how much of the previous two weeks they experienced each feeling on a 6-point scale. Similarly, the European Social Survey has 15 questions on the respondent's affective state over the past week, with responses on a 4-point scale. The SF-36 health measurement tool contains a set of nine items relating directly to the respondent's affective state over the previous four weeks, also using a 6-point scale (Ware and Gandek, 1998). Five of these nine items have been included in the EU-SILC 2013 well-being module to capture affect. However, the length of the reference period in both the EQLS and SF-36 questions is potentially problematic, as errors are likely to increase with the length of the reference period.

While the four-week period used in the SF-36 is well suited for its intended purpose – assessing mental health – recall of affective states is likely to be better when the recall period is short and the question refers to a specific day (see Chapter 2). The affect questions presented in Module C, therefore, are similar in structure to those contained in the

Gallup World Poll (although also drawing on the ESS questionnaire). These questions focus specifically on the affective state of the individual on the previous day. In addition, they ask for a 0 to 10 frequency judgement, rather than requiring judgements relating to the intensity of the feeling.

The affect module includes 10 questions, largely drawn from those used in the Gallup World Poll and those used by the ONS. Four of the questions are related to positive affect and six to negative affect, reflecting the apparent potential for multi-dimensionality in negative affect in particular.

D. Eudaimonic well-being

Eudaimonic well-being encompasses a range of concepts, many with clear policy relevance. Some aspects of eudaimonic well-being – such as meaning or a sense of purpose in life, and a sense of belonging – capture elements of subjective well-being not reflected in life evaluation or affect, but with high intuitive relevance. Other aspects of eudaimonic well-being – such as "agency" and locus of control – are more tenuously related to subjective well-being conceived of as an outcome, but are powerful explanatory variables for other behaviour. However, this level of potential relevance is not matched by an equally good understanding of what eudaimonic well-being actually "is", and more specifically, how it should be measured. In particular, as discussed in Chapter 1, it is not clear whether eudaimonic well-being captures a single underlying construct like life evaluation or is rather an intrinsically multi-dimensional concept like affect. For this reason, the proposed measures of eudaimonic well-being should be considered experimental.

The eudaimonic well-being module proposed here is based on elements of the European Social Survey well-being module and the Flourishing Scale proposed by Diener et al. (2010). This provides a starting point for data producers who wish to collect measures of eudaimonic well-being. The proposed questions are consistent with what is known about best practice in collecting such information (see Chapter 2), but cannot be considered definitive in the absence of a coherent body of international and large, representative-sample research comparable to that existing for life evaluation and affect.

E. Domain evaluations

In addition to evaluating life as a whole, it is also possible to collect information evaluating specific life "domains" such as health or standard of living. Such information has a wide range of potential uses (Dolan and White, 2007; Ravallion, 2012) and may be better adapted to some policy and research questions than over-arching evaluations relating to life as a whole. A challenge to providing advice on measuring domain evaluations is the sheer range of possible life domains that could be measured. Some areas, such as job satisfaction, have substantial literatures in their own right, while others do not. The goal of the domain evaluation module is not to provide an exhaustive approach to measuring subjective evaluations of all policy-relevant life domains. Instead, the focus is on the more limited objective of detailing a limited set of domain evaluation measures that can be used in a general social survey focused on measuring well-being across multiple domains or as the basis for an analysis of the relationship between overall life satisfaction and domain evaluations (e.g. Van Praag, Frijters and Ferrer-i-Carbonell, 2003). Each individual question can, of course, be used in its own right in analysis and monitoring of the particular outcome that it reflects.

Ideally, the questions comprising the domain satisfaction block would meet two key criteria. First, they would be independently meaningful as measures of satisfaction with a particular aspect of life; and second they would collectively cover all significant life domains. A major practical challenge to this sort of approach, however, is that there is no generally agreed framework for identifying how to divide well-being as a whole into different life domains. Different authors have taken different approaches. For example, the Stiglitz/Sen/Fitoussi commission identified eight domains of quality of life alongside economic resources, while the ONS proposal for measuring national well-being identified a slightly different set of nine domains of well-being, including "individual well-being" as a distinct domain capturing overall life evaluations. The OECD's *Better Life Initiative* uses eleven life domains, including a distinct domain on subjective well-being, while the New Zealand General Social Survey uses a slightly different set of ten domains. Another approach is that adopted for constructing the *Personal Wellbeing Index* (PWI; International Wellbeing Group, 2006). This consists of eight primary items that are meaningful on their own but which can also be used to calculate an overall index of subjective well-being. The domains that are included in the PWI have been subject to considerable testing and reflect the results of extensive factor analysis. Table 3.1 compares these different approaches.

Table 3.1. **A comparison of life domains**

SSF	BLI	ONS	NZGSS	PWI
Material conditions	Income and wealth	Personal finance	Economic standard of living	Standard of living
Economic insecurity		The economy		Future security
	Housing			
	Jobs and earnings	What we do	Paid work	
Personal activities	Work and life balance		Leisure and recreation	
Health	Health status	Health (physical and mental)	Health	Personal health
Education	Education and skills	Education and skills	Knowledge and skills	
Social connections	Social connections	Our relationships	Social connectedness	Personal relationships
				Community connectedness
Political voice and governance	Civic engagement and governance	Governance	Civil and political rights	
Personal insecurity	Personal security	Where we live	Safety	Personal safety
Environmental conditions	Environmental quality	The environment	The environment	
			Cultural identity	
				Achieving in life
	Subjective well-being	Individual well-being	Life satisfaction	

Note: The SSF domains have been widely used by other agencies as the basis of their own analysis of well-being. The European Union Sponsorship Group on Measuring Progress, Well-being and Sustainable Development, the UNECD Sustainable Development Task Force, and the ISTAT, for example, have all adopted variants of the SSF approach for various purposes.

Source: The acronyms used in Table 3.1 are: *Commission on the Measurement of Economic Performance and Social Progress* – Sen, Stiglitz, Fitoussi (SSF); *Your Better Life Index* – OECD (BLI); Office for National Statistics (ONS); New Zealand General Social Survey (NZGSS); and *Personal Wellbeing Index* (PWI).

There is a high degree of overlap in the different approaches outlined in Table 3.1. The proposed domain evaluations module draws on these to identify ten questions pertaining to ten specific life domains. These domains include the constituent elements of the PWI as a subset, but also three additional domains (time to do what you like doing, the quality of the environment, and your job) that are of potential policy relevance in and of themselves. The nine domains are:

- Standard of living.

- Health status.
- Achievement in life.
- Personal relationships.
- Personal safety.
- Feeling part of a community.
- Future security.
- Time to do what you like doing.
- Quality of the environment.
- Your job (for the employed).

The ten proposed questions cover all of the main domains of well-being identified in Table 3.1 except one: governance. The range of concepts covered by political voice, governance and civil and political rights is very broad, and there is no model question or set of questions that could be used as the basis for inclusion in these guidelines. Similarly, there would be little value in developing a question from scratch without testing to see how the question performs. However, governance is undeniably an important dimension of well-being. The issue of how best to collect information on satisfaction with governance, political voice and civil and political rights therefore remains a key area for future research.

F. Experienced well-being

As noted in the section describing Module C (affect), the gold standard for measuring affect is via the experience sampling method. When this is not possible, the day reconstruction method (DRM) provides a well-tested methodology that produces results consistent with the experience sampling method (Kahneman and Krueger, 2006). Although it is not possible to implement the DRM in general household surveys, it is possible in time-use surveys. This module provides approaches to implementing the measurement of affect in time-use diaries. Because of the value that time-use diary information on subjective well-being adds, and because information on affect yesterday from general household surveys is not a good substitute[15] for measures like those collected through the DRM, it is strongly recommended that information on experienced well-being be collected in time-use surveys whenever possible.

The experienced well-being module presents two approaches to measuring subjective well-being in time-use diaries. The first is essentially the implementation of the DRM used in the 2011 American Time Use Survey (ATUS). This provides aggregate information similar to the full DRM, but restricts the information collected to only three diary episodes per respondent. This helps reduce the respondent burden and the amount of interviewer time required per respondent, which is otherwise relatively high with the full DRM. The data collected using this method is exceptionally rich, as it involves collecting information on a number of different moods and feelings. As with the affect questions in Module C, it uses a 0-10 scale. This is longer than the 0-6 scale currently used in the ATUS, but is preferred for reasons of consistency with other scales used in these guidelines and because the (relatively limited) literature on the subject tends to support the choice of the longer scale (Kroh, 2006; Cummins and Gullone, 2000).

An alternative to the DRM is also included in the experienced well-being module. This is based on the "unpleasant/pleasant" ("très désagréable/très agréable") approach used by the INSEE in the *Enquête Emploi du temps 2010*. Although the INSEE approach captures less

information than the DRM – the measure used to collect information on affective state is uni-dimensional – it does have two significant advantages. First, it is a self-complete question that can be included on the diary form. This significantly reduces interviewer time and the associated costs, and does not add much to the time required for respondents to fill in the diary (INSEE, 2010). Because of this, information can be collected on the respondent's affective state during all diary episodes, allowing more comprehensive analysis. The self-completed nature of the question also makes it potentially suitable for inclusion in "light" time-use surveys that rely more heavily on respondents to self-complete their diary. The second point in favour of the INSEE approach is that analysis of the available data suggests that the results are broadly comparable with results from the DRM when these are reduced to a uni-dimensional measure such as the "U-index" or affect balance.

There is currently relatively little basis to assess which method is preferable overall. The DRM is better grounded in the research literature, with good evidence of its validity, and provides a more detailed view of the different moods people experience. On the other hand, the INSEE approach appears to manage adequate data quality combined with significantly lower respondent and interviewer burden, as well as detail on a complete sample of episodes. Resolving the issue of which approach is to be preferred will require further analysis, drawing on data derived from both methodologies. For this reason, both approaches are detailed in the experienced well-being module.

Question templates

The six question modules are attached to these guidelines as Annex B. Each question module is presented in the same format, containing a common set of headings that outline the objectives of the module (what kind of information it is trying to gather), a description of the contents of the module, the origin of the questions in the module, how the data from the module should be presented, background information for interviewers, and the detailed question wording. These headings are described in more detail below.

Objectives

The objective succinctly outlines the purpose of the block, including both the type of information it is designed to elicit and the rationale behind the scope of the question block.

Description

A description of the contents of each question module is provided, outlining the role of each of the questions in the module with respect to the module's objectives. The description is intended to assist users to identify which questions they wish to use in the event that they choose to implement only part of the module.

Origin

Questions included in each module are drawn from existing sources and remain unchanged wherever possible to maximise comparability with previous work. However, some items have been modified to a greater or lesser extent where a variety of question versions exist in the literature, and/or where there are clear grounds for small changes in item wording or response scales, for example based on the evidence in Chapter 2. The origin section indicates the source of the questions and notes whether they have been altered.

Completion time

This gives an estimate of the time required to run the entire module.

Output

The output section contains basic information on the production of standard tables and measures from the question block. This information is not exhaustive, but is intended to provide some basic guidelines for data producers. Such guidelines are important, both in order to assist producers in presenting the data appropriately, but also to provide context for why the questions are framed in the way that they are.

A number of the question blocks are intended to produce multi-item measures of subjective well-being derived from the survey questions. The output section provides details on the construction of these multi-item measures, and how they should be reported.

Guidelines

The quality of any survey data is heavily influenced by the attitude of the respondents to the questions they are being asked. Although the evidence is overwhelming that measures of subjective well-being are not regarded as particularly challenging or awkward by respondents (particularly when compared to questions on some other commonly-asked topics, such as income), better-quality information is likely to result if interviewers understand what information is being collected and how it will be used, and they are able to communicate this clearly to respondents. This enables interviewers to answer queries from respondents on why the information is important or on what concept the question is trying to elicit from them.

The guidelines for interviewers contained in this module are not intended as a substitute for the more extensive notes and/or training that would normally be provided to interviewers in the process of preparing to conduct a household survey. However, they do provide a basis from which users of the module can develop their own more substantive guidelines.

4. Survey implementation

How a survey is implemented is crucial to its effectiveness. A carelessly-implemented survey will result in low-quality and unreliable data regardless of the quality of the underlying questionnaire. In general, the features relevant to the effective implementation of any household survey also hold for those collecting information on subjective well-being. These guidelines make no attempt to provide a detailed discussion of best practice in survey implementation, for which high-quality standards and guidelines already exist (UN, 1984). However, there are several points where the specific nature of measures of subjective well-being raises survey implementation issues that are worth noting.

Interviewer training

Interviewer training is crucial to the quality of responses in any survey. However, the measurement of subjective well-being raises additional issues because the subject matter may be unfamiliar to interviewers. This is, ironically, particularly so for national statistical agencies with a permanent force of field interviewers. Although a body of trained interviewers will generally contribute to higher response rates and better responses, interviewers may struggle with questions if they cannot explain adequately to respondents why collecting such information is important and how it will be used. Anecdotal evidence

and feedback from cognitive testing shows that this can be an issue with some subjective measures, particularly measures of affect (ONS, 2012). In some cases, respondents may find it difficult to understand why government might want to collect this information and that the concept that the survey intends to collect is their recently-experienced affective state rather than their normal affective state.

To manage risks around respondent attitudes to questions on subjective well-being, it is imperative that interviewers are well-briefed, not just on what concepts the questions are trying to measure, but also on how the information collected will be used. This is essential for interviewers to build a rapport with respondents and can be expected to improve compliance by respondents and the quality of responses. While the notes on interviewer guidelines contained in the question modules provide some crucial information specific to each set of questions, a more comprehensive approach should draw on information on the validity and use of measures of subjective well-being (Chapter 1) and the analysis of subjective well-being data (Chapter 4).

Ethical issues

Evidence suggests that measures of subjective well-being are relatively non-problematic for respondents to answer. Rates of refusal to respond are low, both for life evaluations and for measures of affect (Smith, 2013). In general, item-specific non-response rates for subjective well-being measures are similar to those for marital status, education and labour market status, and much lower than for measures of income (Smith, 2013). This suggests that, in general, such questions are not perceived as problematic by respondents.

Cognitive testing of measures of subjective well-being supports the conclusions reached from an examination of item-specific non-response rates (ONS, 2012), with some notable exceptions. In particular, the ONS found that eudaimonic questions relating to whether respondents felt that what they did in life were worthwhile and the experience of loneliness caused visible distress in some respondents, particularly among disabled and unemployed respondents.

Best practice suggests that statistical providers should consider how to manage the risks associated with questions that are distressing to respondents. Although it is important not to overstate the risks – they apply mainly to eudaimonic questions, and to a small proportion of respondents – such issues should be dealt with effectively. A complicating factor is that it might not be evident at the time of the interview whether a respondent has been affected by the questioning. One approach to managing this proposed by the ONS (2012) is to distribute a leaflet at the time of the interview giving respondents information on the purpose of the survey and reiterating the confidentiality of the data collected. The leaflet would also contain information for distressed respondents about where to seek help.

Coding and data processing

The coding of information on subjective well-being is generally straight-forward. In general, numerical scales should be coded as numbers, even if the scale bounds have labels. Much analysis of subjective well-being data is likely to be quantitative and will involve manipulating the data as if it were cardinal. Even for fully-labelled response scales (such as the "yes/no" responses that apply to many questions), it is good practice to code the data numerically as well as in a labelled format in order to facilitate use of the micro-data to

produce summary measures of affect balance or similar indices. "Don't know" and "refused to answer" responses should be coded separately from each other as the differences between them are of methodological interest.

Normal data-cleaning procedures include looking for obvious errors such as data coders transposing numbers, duplicate records, loss of records, incomplete responses, out-of-range responses or failure to follow correct skip patterns. Some issues are of particular relevance to subjective data. In particular, where a module comprising several questions with the same scale is used, data cleaning should also involve checking for response sets (see Chapter 2). Response sets occur when a respondent provides identical ratings to a series of different items. For example, a respondent may answer "0" to all ten domain evaluation questions from Module E. This typically suggests that the respondent is not, in fact, responding meaningfully to the question and is simply moving through the questionnaire as rapidly as possible. Such responses should be treated as a non-response and discarded. In addition, interviewer comments provide an opportunity to identify whether the respondent was responding correctly, and a robust survey process will make provision for allowing such responses to be flagged without wiping the data record.

Finally, it is important to emphasise that much of the value from collecting measures of subjective well-being comes from micro-data analysis. In particular, analysis of the joint distribution of subjective well-being and other outcomes and use of subjective well-being measures in cost-benefit analysis cannot usually be accomplished through secondary use of tables of aggregate data. Because of this, a clear and comprehensive data dictionary should be regarded as an essential output in any project focusing on subjective well-being. This data dictionary should have information on survey methodology, sampling frame and correct application of survey weights, as well as a description of each variable (covering the variable name, the question used to collect it and how the data is coded). If a variable is collected from only part of the survey sample due to question routing, this should also be clearly noted in the data dictionary.

Notes

1. See the section of Chapter 4 (*Output and analysis of subjective well-being measures*) relating to cost-benefit analysis for an example of this distinction when comparing stated preference approaches to estimating non-market values as opposed to using subjective well-being measures.
2. The distinction between cardinal and ordinal measures is important to measuring subjective well-being. With ordinal measures the responses are assumed to show the rank order of different states, but not the magnitude. For example, with ordinal data a 5 is considered higher than a 4 and an 8 is considered higher than a 7. However, nothing can be said about the relative size of the differences implied by different responses. For cardinal data it is assumed that the absolute magnitude of the response is meaningful, and that each scale step represents the same amount. Thus, a person with a life satisfaction of 5 would be more satisfied than someone reporting a 4 by the same amount as someone reporting an 8 compared to a 7. Most subjective well-being measures are technically ordinal, but the evidence suggests that treating them as cardinal does not generally bias the results obtained (Ferrer-i-Carbonell and Frijters, 2004).
3. In many cases it may be possible to collect one income measure (say, gross household income) and impute net household income on the basis of household size and composition and the relevant tax rates and transfer eligibility rules.
4. One example of these sorts of measure is provided by the material deprivation questions contained in the EU-SILC.
5. See for example the Eurostat social inclusion and living conditions database, *http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/data/database*.

6. These other factors potentially include both the confounding effects of shared method variance if the quality-of-life measures in question are subjective, or more substantive factors such as events earlier in the life course that may impact both income and quality of life.
7. The Five Factor Model is a psychological framework for analysing personality type. It identifies five main factors that relate to personality: Neuroticism; Extraversion; Openness to experience; Agreeableness; and Conscientiousness. The scale is widely used in psychological research and is well suited to inclusion in survey questionnaires.
8. "Frame of reference" refers to the situation or group on which respondent's base comparisons when formulating a judgement about their lives or feelings. The respondent's knowledge of how others live and their own prior experiences can influence the basis on which judgements are reached about the respondent's current status.
9. This is not, in fact, beyond the realm of possibility. Many government agencies may have an interest in collecting measures of client satisfaction. However, the case for collecting general measures of subjective well-being as a standard part of interactions with government service delivery agencies is beyond the scope of this paper.
10. The need for a relatively large sample size is one reason to prefer a simple measure of subjective well-being with a low respondent burden in place of a technically more reliable multi-item measure with a higher respondent burden. The quality gains from a more detailed measure need to be assessed carefully against the quality losses associated with any reduction in sample size associated with a longer measure.
11. Internet surveys are, from this perspective, a way of implementing CASI.
12. In this case the precise transition question used was: "Now thinking about your personal life, are you satisfied with your personal life today", and the subjective well-being measure that followed was the Cantril self-anchoring ladder of life measure. It does not follow that the same transition question will work in other contexts, and transition questions should be tested empirically before being relied on.
13. Some versions of the satisfaction with life question use different response scales, such as a 5-point labelled Likert scale or a 1-10 scale. Based on the conclusions from Chapter 2, the core module uses a 0-10 end-labelled scale.
14. Technically the Circumplex model implies that positive and negative affect are ends of a single dimension rather than a way of grouping several independent types of feeling. Here the Circumplex model is used as an organising framework to help impose some structure on the range of different affective states, without assuming continuity on the positive/negative axis.
15. Information on affect yesterday from general household surveys, while of interest in its own right, does not allow analysis of how different activities, locations and the people with whom the respondent is with impact on subjective well-being.

Bibliography

Balestra, C. and J. Sultan (2013), "Home Sweet Home: The Determinants of Residential Satisfaction and Its Relation with Well-Being", *OECD Statistics Directorate Working Papers* (forthcoming), OECD, Paris.

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

Blanchflower, D. and A. Oswald (2011), "International happiness", *NBER Working Paper*, No. 16668, National Bureau of Economic Research.

Blanchflower, D. and A. Oswald (2008), "Is well-being U-shaped over the life cycle?", *Social Science and Medicine*, Vol. 66(8).

Boarini, R., M. Comola, C. Smith, R. Manchin and F. De Keulenaer (2012), *What Makes for a Better Life? The determinants of subjective well-being in OECD countries: Evidence from the Gallup World Poll*, STD/DOC(2012)3, OECD.

Card, D., A. Mas, E. Moretti and E. Saez (2010), "Inequality at work: the effect of peer salaries on job satisfaction", *NBER Working Paper*, No. 16396, National Bureau of Economic Research.

Clark, A.E., Y. Georgellis and P. Sanfey (1998), "Job Satisfaction, Wage Changes and Quits: Evidence From Germany", *Research in Labor Economics*, Vol. 17.

Costa, P.T. Jr. and R.R. McCrae (1992), *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual*, Odessa, FL, Psychological Assessment Resources.

Cummins, R.A., R. Eckerslet, J. Pallant, J. Vugt and R. Misajon (2003), "Developing a National Index of Subjective Wellbeing: The Australian Unity Wellbeing Index", *Social Indicators Research*, No. 64, pp. 159-190.

Cummins, R.A. and E. Gullone (2000), "Why we should not use 5-point Likert scales: The case for subjective quality of life measurement", *Proceedings, Second International Conference on Quality of Life in Cities*, National University of Singapore, pp. 74-93.

Davern, M., R. Cummins and M. Stokes (2007), "Subjective Wellbeing as an Affective-Cognitive Construct", *Journal of Happiness Studies*, No. 8, pp. 429-449.

Deaton, A. (2011), "The financial crisis and the well-being of Americans", *Oxford Economic Papers*, No. 64, pp. 1-26.

Diener, E., J. Helliwell and D. Kahneman (eds.) (2010), *International Differences in Well-Being*, Oxford University Press.

Diener, E., S. Oishi and R. Lucas (2003), "Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life", *Annual Review of Psychology*, No. 54, pp. 403-425.

Dolan, P., T. Peasgood and M. White (2008), "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being", *Journal of Economic Psychology*, Vol. 29, pp. 94-122.

Dolan, P. and M. White (2007), "How can measures of subjective well-being be used to inform policy?", *Perspectives on Psychological Science*, Vol. 2(1), pp. 71-85.

Easterlin, R. (1974), "Does Economic Growth Improve the Human Lot? Some Empirical Evidence", in David, P.A. and M. W.Reder, *Nations and Households in Economic Growth: Essays in Honour of Moses Abramovitz*, New York, Academic Press Inc, pp. 89-125.

Eurofound (2007), *2007 European Quality of Life Survey Questionnaire*.

European Social Survey (2007), *Final Source Questionnaire Amendment 03*.

Eurostat (2005), *Guidelines for the development and criteria for the adoption of Health Survey instruments*, Eurostat, Luxembourg.

Eurostat (2004), *Guidelines on harmonised European Time Use Surveys*.

Ferrer-i-Carbonell, A. and P. Frijters (2004), "How important is methodology for the estimates of the determinants of happiness?", *The Economic Journal*, No. 114, pp. 641-659.

Frey, B.S. and A. Stutzer (2000), "Happiness, Economy and Institutions", *The Economic Journal*, Vol. 110(466), pp. 918-938.

Frey, B.S. and A. Stutzer (2008), "Stress that Doesn't Pay: The Commuting Paradox", *Scandinavian Journal of Economics*, Vol. 110(2), pp. 339-366.

Gallup Organisation (2012), *Indexes and Questions*.

Goldberg, D.P. et al. (1978), *Manual of the General Health Questionnaire*, Windsor, England, NFER Publishing.

Gutiérez, J., B. Jiménez, E. Hernández and C. Puente (2005), "Personality and subjective well-being: big five correlates and demographic variables", *Personality and Individual Differences*, No. 38, pp. 1561-1569.

Harkness, J., B.E. Pennell and A. Schoua-Glusberg (2004), "Survey Questionnaire Translation and Assessment", in S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, John Wiley and Sons, Inc, Hoboken, NJ, USA.

Helliwell, J.F. (2008), "Life Satisfaction and the Quality of Development", *NBER Working Paper*, No. 14507, National Bureau of Economic Research.

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

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

Helliwell, J.F. and S. Wang (2011b), "Trust and Well-being", *International Journal of Wellbeing*, available online at: *www.internationaljournalofwellbeing.org/index.php/ijow/issue/current*.

Huppert, F and T. So (2008), "Deriving an objective definition of well-being", *Working Paper,* Well-being Institute, University of Cambridge. See also J. Michaelson, S. Abdallah, N. Steur, S. Thompson and N. Marks, *National Accounts of Well-being: Bringing real wealth onto the balance sheet*, New Economics Foundation.

INSEE (2010), *Enquête Emploi du temps*.

International Wellbeing Group (2006), *Personal Wellbeing Index*, 4th edition, Melbourne, Australian Centre on Quality of Life, Deakin University, available online at: *www.deakin.edu.au/research/acqol/instruments/wellbeing_index.htm*.

Kahneman, D. and A. Deaton (2010), "High income improves life evaluation but not emotional well-being", *Proceedings of the National Academy of Sciences*, Vol. 107(38), pp. 16489-16493.

Kahneman, D., E. Diener and N. Schwarz (1999), *Well-being. The Foundations of Hedonic Psychology*, Russel Sage Foundation, New York.

Kahneman, D. and A.B. Krueger (2006), "Developments in the Measurement of Subjective Well-Being", *Journal of Economic Perspectives*, Vol. 20(1), pp. 19-20.

Kroh, M. (2006), "An experimental evaluation of popular well-being measures", *DIW Berlin Working Paper*, No. 546.

Krueger, A.B. and A.I. Mueller (2012), "Time Use, Emotional Well-Being, and Unemployment: Evidence from Longitudinal Data", *The American Economic Review*, Vol. 102(3), pp. 594-599.

Larsen, R.J. and B.L. Fredrickson (1999), "Measurement issues in emotion research. Well-being", *The Foundations of Hedonic Psychology*, pp. 40-60.

Lucas, R.E. (2007). "Long-term disability is associated with lasting changes in subjective well-being: evidence from two nationally representative longitudinal studies", *Journal of Personality and Social Psychology*, Vol. 92(4), p. 717.

Lucas, R.E., A. Clark, Y. Georgellis and E. Diener (2004), "Unemployment alters the set point for life satisfaction", *Psychological Science*, No. 15, pp. 8-13.

NEF (2012), *Well-being patterns uncovered: An analysis of UK data*, United Kingdom.

OECD (2009), *Doing Better for Children*, OECD Publishing, Paris.

ONS (2012), *Subjective Well-being: A qualitative investigation of subjective well-being questions*, ONS, United Kingdom.

ONS (2011), *Initial investigations into Subjective Well-being from the Opinions Survey*, ONS, United Kingdom.

Oswald, F., H. Wahl, H. Mollenkopf and O. Schilling (2003), "Housing and Life Satisfaction of Older Adults in Two Rural Regions in Germany", *Research on Ageing*, Vol. 25(2), pp. 122-143.

Pavot, W. and E. Diener (1993), "Review of the Satisfaction With Life Scale", *Psychological Assessment*, Vol. 5(2), pp. 164-172.

Ravallion, M. (2012), "Poor, or just feeling poor? On using subjective data in measuring poverty", *World Bank Policy Research Working Paper*, No. 5968, World Bank Development Research Group, Washington, DC.

Russell, J. (1980), "A Circumplex Model of Affect", *Journal of Personality and Social Psychology*, Vol. 39(6), pp. 1161-1178.

Sacks, W.D., B. Stevenson and J. Wolfers (2010), "Subjective Well-being, Income, Economic Development and Growth", *NBER Working Paper*, No. 16441.

Silva, J., F. De Keulenaer and N. Johnstone (2012), "Individual and Contextual Determinants of Satisfaction with Air Quality and Subjective Well-Being: Evidence based on Micro-Data", *OECD Environment Directorate Working Paper*, OECD Publishing, Paris.

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.

Stiglitz, J.E., A. Sen and J.P. Fitoussi (2009), *Report by the Commission on the Measurement of Economic Performance and Social Progress*.

Tinkler, L. and S. Hicks (2011), *Measuring Subjective Well-being*, ONS, United Kingdom.

UNECE Secretariat (2009), *Revised terms of reference of UNECE/WHO/Eurostat steering group and task force on measuring health status*, UNECE.

UNECE (2010), *Manual on Victimization Surveys*, United Nations.

UNICEF (2007), *Child poverty in perspective: An overview of child well-being in rich countries*, Innocenti Report Card 7.

United Nations Statistical Division (1986), *National Household Survey Capability Programme, Sampling Frames and Sample Designs for Integrated Survey Programmes. Preliminary version*, United Nations, New York.

United Nations Statistical Division (1984), *Handbook of Household Surveys*, United Nations, New York.

Van Praag, B.M.S., P. Frijters and A. Ferrer-i-Carbonell (2003), "The anatomy of subjective well-being", *Journal of Economic Behaviour and Organisation*, No. 51, pp. 29-49.

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.

Ware, J. and B. Gandek (1998), "Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project", *Journal of Clinical Epidemiology*, Vol. 51(11), pp. 903-912.

Weinhold, D. (2008), "How big a problem is noise pollution? A brief happiness analysis by a perturbable economist", *MPRA Working Paper*, No. 10660.

Winkelmann, L. and R. Winkelmann (1998), "Why Are The Unemployed So Unhappy? Evidence From Panel Data?", *Economica*, Vol. 65, pp. 1-15.

World Health Organisation (2012), *World Health Survey Instruments and Related Documents*.

Chapter 4

Output and analysis of subjective well-being measures

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Introduction

This chapter provides guidance regarding the release and use of subjective well-being data. It briefly re-caps the policy and public interest in the data (outlined in Chapter 1, *Concept and validity*), before covering how information can be reported and analysed. This includes the statistical outputs that may be released; basic information about the methods of analysis that may be adopted; and a discussion of key interpretive issues, placing particular emphasis on the extent to which levels of subjective well-being can be expected to vary in different circumstances.

The chapter is divided into three main sections, summarised in Table 4.1. The first and largest section focuses on the use of subjective well-being data to complement existing measures of well-being. This includes examination of trends in subjective well-being over time, the distribution of subjective well-being across different groups within society, and its distribution across different countries. The first part of this section outlines approaches to measuring well-being and the value added that subjective well-being brings relative to other measures. The second part of the section then describes how subjective well-being can be reported, including the summary statistics that may be of interest. Finally, issues in the analysis and interpretation of descriptive statistics on subjective well-being are explored. These include the size of change over time, or difference between groups, that can be expected, as well as the risk of cultural "bias" in cross-country comparisons.

The remaining two sections of the chapter deal with more detailed analyses of subjective well-being, which might be conducted by government analysts and others on the basis of micro-data released by statistical agencies. Section 2 addresses analyses of the *drivers* of subjective well-being. This includes the relationship between subjective well-being and other important well-being outcomes, such as income and health, as well as the use of subjective well-being data to inform the appraisal, design and evaluation of policy options. Section 3 addresses subjective well-being data as an input for other analyses. First, it considers the use of subjective well-being as an explanatory variable for other outcomes, and then focuses on the potential use of subjective well-being data in cost-benefit analysis.

Section 1 will be of most direct interest to large-scale data producers, such as national statistical agencies, as it concerns the kinds of outputs and analyses that they are most likely to report for a wide range of audiences. Sections 2 and 3 provide a sense of the broader uses of subjective well-being data – which are essential to consider when planning its measurement (as set out in Chapter 3, an approach to *Measuring subjective well-being*). Analyses of drivers, for example, require consideration of the co-variates to be collected alongside subjective well-being data, and ideally call for data from which causal inferences can be drawn. The potential risk of measurement error, and the various biases that may be present in the data, are also major themes throughout the chapter. However, as the relevance of measurement errors depends on the *intended usage* of the data (Frey and Stutzer, 2002), the chapter is organised around data uses, rather than around these sources of error. Key interpretive issues for each type of analysis are summarised in Table 4.1.

Table 4.1. **Summarising possible uses of subjective well-being data**

Data use	What	Why	Who	Key interpretive issues
1) Complementing existing measures of well-being	Core measures/headline indicators used to examine: *i)* National trends over time. *ii)* Distribution of outcomes across different groups within society. *iii)* Distribution of outcomes across countries. Includes indicators of central tendency or "level", as well as distribution, and the relative *rate* of rise or decline over time.	To know if the changes affecting society have an impact on subjective well-being. To identify vulnerable groups and areas of suffering – highlighting where key drivers of subjective well-being may lie – and where there may be opportunities for policy interventions. To conduct international benchmarking, assist in the interpretation of national data, and identify where countries may be able to learn from others' experiences.	Governments (central, regional, local). Wider public. Public, private and third sector organisations. Researchers interested in country-level drivers of national well-being. Individuals and organisations – e.g. making decisions about where to live and work.	*i)* What size of difference between groups or over time can be expected? *ii)* What alternative explanations should be considered for observed differences? *iii)* What is the role of culture and cultural bias in cross-country comparisons?
2) Better understanding the drivers of subjective well-being	Analyses based on national and international micro-data, with subjective well-being used as the dependent variable, to: *i)* Examine the relationship between subjective well-being and other important life circumstances, such as income and health. *ii)* Inform policy options appraisal, design and evaluation. *iii)* Inform policy trade-offs.	To improve our understanding of well-being overall, by examining the relationship between subjective well-being, life circumstances, and other important well-being outcomes. To highlight areas of policy with the greatest potential to improve subjective well-being, and the life events/circumstances most likely to put subjective well-being at risk. To assist in government decision-making processes, including the allocation of resources and the design elements of policies. To inform the public and employers about the likely drivers of individual subjective well-being, providing better information for individual and organisational decision-making.	Governments. Researchers. Individuals wanting better information to support decision-making. Employers wanting to understand and improve employee well-being.	*i)* What size of impact can be expected? *ii)* How can the impacts of different drivers be compared?
3) Subjective well-being as an input for other analyses, particularly cost-benefit analysis	Micro-data on subjective well-being, used as an input for other analyses, including: *i)* As an explanatory variable for other elements of well-being or behaviour. *ii)* Used to estimate the value of non-market goods and services, for the purposes of cost-benefit analyse.s	To better understand how subjective well-being can contribute to other well-being outcomes and shed light on human decision-making processes, including the various biases that may be present. To provide an alternative to traditional economic approaches to estimating the value of non-market goods, supporting government (and other organisations) in making decisions about complex social choices.	Researchers. Governments. Individuals wanting better information to support decision-making. Employers wanting to understand and improve employee well-being.	*i)* The sensitivity of subjective well-being data to non-market goods. *ii)* Measurement error and its impact on valuations. *iii)* Co-variates to include in regression models. *iv)* Time horizons for study.

1. Using subjective well-being to complement other outcome measures

Introduction

Subjective well-being is an essential element of a broader and multi-dimensional concept of human well-being and can be used in the context of monitoring reports on the living conditions of different countries or sub-national units. Indicators of interest will include the overall level of subjective well-being, its rate of change over time and its distribution across different groups within society. This section is organised in three parts. The first part addresses what is meant by "measuring well-being" and briefly discusses how subjective well-being can contribute in this area. This includes outlining what subjective well-being data can add to more conventional measures and why subjective well-being might be considered an important outcome in its own right. The second part focuses on reporting measures of subjective well-being. This examines the relative merits of a number of different approaches to summarising and describing subjective well-being

data. The section concludes with discussion of the issues arising in analyses that aim to compare different levels of subjective well-being. This includes consideration of how to interpret observed differences between groups, over time and between different countries – including when there may be a risk of cultural "bias" in the data.

What does "measuring well-being?" mean, and why do it?

Measuring human well-being involves identifying the key components of a good life and then selecting a set of indicators that provide information about the progress of society with respect to these outcomes. There are three key elements of well-being that it will be important to measure: i) trends over time; ii) the distribution of outcomes across different members of society; and iii) the distribution of outcomes across countries.

Measures of well-being are important to governments and to the general public. Although many societal outcomes are often not under direct government control, governments still seek to have a positive influence on well-being and are often called to intervene to address poor outcomes and/or declines in well-being over time. Governments will generally have an interest in all three elements: trends over time; distributions across society; and international benchmarking.

Businesses and voluntary sector organisations may also have an interest in monitoring well-being. National trends influence the business environment. Businesses can play a role in meeting the needs of vulnerable groups and bridging inequalities in society, and they may look to international measures when considering export, expansion and/or relocation opportunities. Voluntary sector organisations may have a strong interest in the distribution of outcomes across society – including what this tells us about vulnerable groups and the support that they may need. Voluntary sector organisations may also look overseas for examples of different practices, and some voluntary organisations will be engaged in international work that seeks to address global inequalities in well-being outcomes.

Approaches to measuring well-being

There are a number of different approaches to monitoring well-being through dedicated reports. GDP per capita is commonly used as a proxy measure for the overall well-being of countries. Other commonly-cited indicators of national progress include poverty, unemployment levels, infant mortality, life expectancy, educational attainment, crime figures and air quality. These provide information on outcomes that may not be accurately captured by GDP per capita, but which are important to well-being.

Nonetheless, it can be difficult to compile a coherent overall picture of well-being from a disparate range of measures. One approach is to develop composite indices, such as in the UN's Human Development Index (HDI),[1] which combines information on life expectancy at birth, mean years of schooling, expected years of schooling and gross national income per capita, to produce a single overall figure. Alternatively, a range of indicators can be presented in a "dashboard", such as that adopted in *How's Life?* (OECD, 2011a) or the various sets of sustainable development indicators available, such as those in the EU sustainable development strategy (Eurostat, 2009), or *Measuring New Zealand's Progress Using a Sustainable Development Approach* (Statistics New Zealand, 2008).

The role of subjective well-being in measuring well-being

What people feel about their lives matters. A nation of materially wealthy, healthy, but miserable citizens is not the kind of place where most people would want to live. The available evidence suggests that the general public, at least in affluent countries, do regard subjective well-being as an important component of national well-being overall. For example, Dolan and Metcalfe (2011) report an initial survey asking UK respondents to rank seven ways of measuring progress, in which "people's happiness" was ranked behind the "state of the economy" and "peoples' health" , but above "crime rates", "education levels", "the environment" and "depression rates".

A recent public consultation by the UK Office for National Statistics (ONS, 2011a) found that 79% of 6 870 respondents endorsed "life satisfaction" as a measure of "national well-being and how life in the UK is changing over time" – second only to "health statistics" (80%), with measures such as "income distributions" endorsed by 62%, and "economic measures such as GDP" endorsed by just 30% of respondents. The OECD's web-based interactive tool *Your Better Life Index* offers individuals the opportunity to create their own international well-being index, rating the importance of 11 different dimensions of well-being on a 1-5 scale. Ratings shared by around 4 000 users of the website (OECD, 2011b) indicate that life satisfaction is the domain most often ranked the highest (with over 10% of users identifying it as the most important domain), closely followed by health, education, the environment and work-life balance.[2]

One benefit of using subjective well-being to complement existing measures of national progress is that it emphasises the views of individuals. It thus presents an overall picture of well-being that is grounded in people's preferences, rather than in *a priori* judgements about what *should* be the most important aspects of well-being. Subjective well-being measures reflect the unique mix of factors that influence an individual's feelings and assessments. This is not to say that subjective well-being should *replace* other important economic, social and environmental indicators, but it does provide a useful and easy-to-understand complement to existing measures, because it can indicate the *combined impact* of life circumstances on subjective perceptions and emotions.

Subjective well-being measures may also capture some aspects of well-being that are difficult to otherwise observe or quantify through more traditional measures. An example of this, cited in Chapter 1 (Box 1.2), is the marked decline in evaluative measures of subjective well-being in Egypt and Tunisia in the years preceding the 2011 "Arab Spring". Conventional indicators of progress, such as economic growth, and the UN's Human Development Index, continued to rise during this period – thus failing to detect an important social trend.

The public policy applications of subjective well-being measures (described in Chapter 1) are wide-ranging. Extensive reviews on this topic have been published recently by Diener, Lucas, Schimmack and Helliwell (2009), Bok (2010), the European Commission (Chapple et al., 2010), and the New Economics Foundation (Stoll, Michaelson and Seaford, 2012). These reviews build on the earlier conceptual work of Kahneman et al. (2004), Layard (2005), Dolan and White (2007) and Krueger (2009), to name just a few. Specific examples from the field include using life satisfaction and eudaimonic indicators alongside a wide variety of outcome measures to evaluate public projects to enhance well-being, such as the UK Big Lottery Fund well-being evaluation (CLES Consulting and NEF, 2011); and the evaluation of the Community Employment Innovation Project in Canada (Gyarmati et al., 2008), as well as for cost-benefit analyses of psychological therapy (Layard et al., 2007),

estimating the well-being impact of various policy-relevant daily activities, such as commuting (Kahneman and Krueger, 2006; Stutzer and Frey, 2008), as well as to explore policy trade-offs, such as those between inflation and unemployment (Di Tella, MacCulloch and Oswald, 2001) or income and airport noise (Van Praag and Baarsma, 2005). Research linking subjective well-being, and particularly positive affect, to health outcomes (Pressman and Cohen, 2005; Danner, Snowdon and Friesen, 2001; Cohen et al., 2003; Kiecolt-Glaser et al., 2002; and Steptoe, Wardle and Marmot, 2005), as well as income, employment outcomes and productivity (Diener et al., 2002; Wright and Staw, 1999; Keyes 2006; Clark and Oswald, 2002) also suggests a public interest in monitoring such measures.

Like many other measures of well-being, however, subjective well-being data do come with some notable caveats and trade-offs, specifically around data comparability and the risk of measurement error (Ravillion, 2012; see Chapter 2 for a summary). Some of these risks are common to other self-report measures, including the risk of various response biases, and the impact that both question wording and response formats can have on how people answer questions. Frame-of-reference effects[3] and adaptation[4] to life circumstances over time can also potentially influence the levels of subjective well-being observed among different populations and population sub-groups, as well as the nature of the relationship between subjective well-being and its determinants. These issues mean that subjective well-being data, like most self-reported data, need to be interpreted with care and should be used to *complement* rather than replace other indicators of well-being. Interpretive issues are described at length in the sections that follow.

Reporting subjective well-being data

Using subjective well-being data to complement other measures of well-being requires producers of statistical information to regularly collect and release high-quality nationwide data from large and representative samples. Key audiences include policy-makers, public service providers, private businesses and voluntary sector organisations, researchers and the wider public – all of whom may have an interest in whether, where and when conditions in society are improving. For monitoring exercises in particular, it is important that the figures released *mean something* to the general public, as well as to more specialist audiences (New Economics Foundation, 2009).

Many of these audiences will not read statistical releases directly, but rather will rely on how these are reported in a variety of media. It is therefore important to consider how to package the data in a concise yet precise manner to ensure that the necessary information can be easily located and conveyed with accuracy by other parties.

The language used to describe measures is also important. The term "happiness" is often used as convenient shorthand for subjective well-being, in both popular media and parts of the academic literature – not least because *happiness* may be more attention-grabbing and intuitively appealing. The key risk surrounding the term "happiness" is conceptual confusion: whilst the experience of positive emotion (or positive affect) is an important part of subjective well-being, it represents only part of the over-arching concept, and the term "happiness" underplays the evaluative and eudaimonic aspects of subjective well-being as well as the experience of negative affect (pain, sadness, anxiety, etc.), all of which may be of interest to policy-makers.[5] We therefore recommend against describing results only in terms of "happiness", particularly for data releases from national statistics agencies.

Several authors have also shown a tendency to drop the term "subjective" from their reporting, simply describing results in terms of "well-being". This is also a potential source of confusion. Whilst subjective measures of well-being offer an important insight into respondents' views about their own well-being, the OECD regards subjective measures as only one of several measures required to develop a balanced view of well-being overall (OECD 2011a; Stiglitz, Sen and Fitoussi, 2009). This concurs with the outcome of the UK ONS's recent public consultation on what matters for measuring national well-being (ONS, 2011a). For both the ONS and OECD, measuring well-being requires a mix of subjective and objective indicators, and measures across a variety of other dimensions (e.g. education, health, income and wealth, social connections and the environment, to name just a few) are viewed as an essential part of the overall well-being picture.

These considerations mean it will be important, especially when reporting the results of national surveys, to provide a full description of the indicators used – including the underlying constructs of interest, and what they might reflect in addition to "happiness". This could be accompanied by a brief explanation of the rationale for measuring subjective aspects of well-being and their role in complementing (rather than *replacing*) other well-being indicators. Chapter 1 also discusses these issues.

For the purposes of high-level communication about subjective well-being results, particularly with non-specialist audiences, it is desirable to identify a small set of key measures and figures. These guidelines recommend that this set should include one primary measure of life evaluation and its dispersion, as well as a limited number of affect measures if possible (see Chapter 3). Eudaimonia and domain-specific life evaluations may also be of interest, although, as multi-dimensional constructs, they can be more challenging to convey in single headline figures. There are several different ways in which current levels of subjective well-being data can be presented for the purposes of monitoring progress – and the choice of method should ultimately be driven by user need and demand. Recent examples are available from France's National Institute of Statistics and Economic Studies (INSEE – Godefroy, 2011) and the UK's Office for National Statistics (ONS, 2012). Chapter 3 provides recommendations for the basic output associated with the different question modules proposed as part of these guidelines.

Because of the range of possible approaches to presenting and reporting on subjective well-being data, it is useful to consider the issue within some sort of organising framework. At the most general level, the question of how to report subjective well-being data for the purposes of monitoring progress has four elements:

- How to report *central tendency and level*.
- How to report *distribution*.
- Whether and how to *aggregate* responses.
- How to report *change over time* and *differences between groups*.

Reporting central tendency and level

The most fundamental information to report with respect to subjective well-being is the level of the outcome. This can be thought of as addressing the issue of "how high or low is the level of subjective well-being in the population under consideration?". There are three main approaches to describing the level of either single-item or summed multi-item aggregate measures. First, the frequency of responses can be described by category: this involves presenting the proportion of the population that select each response category of

the subjective well-being scale used. Second, the data can be summarised in relation to one or more thresholds. This involves reporting the proportion of the population with a level of subjective well-being above or below a particular threshold level. Finally, the data can be summarised via some measure of central tendency, such as the mean, median or mode. Each of these three approaches has its own strengths and weaknesses.

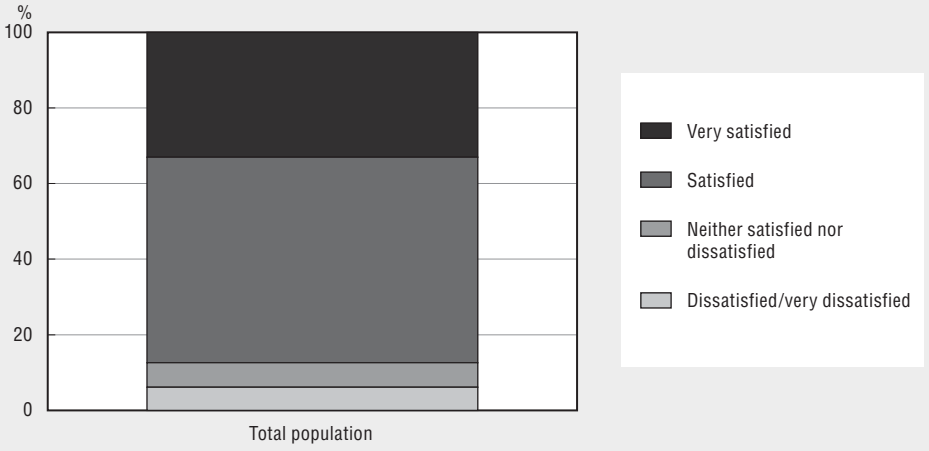
Reporting the proportion of respondents selecting each response category is the method that requires the data producer to make the fewest decisions about presentation. Such an approach has some significant strengths with respect to information on subjective well-being. Because the entire distribution is described, no information is lost. Also, a presentation by category respects the ordinal nature of subjective well-being data[6] and requires no assumptions about the differences among ordinal categories (i.e. there is no assumption that the difference between a 3 and a 4 is the same as that between a 7 and an 8).

However, presenting the whole distribution of responses for each measure also has significant draw-backs. In particular, for a non-specialist audience it is difficult to directly compare two distributions of this sort and reach judgements about which represents a higher or lower "level" of well-being – although non-parametric statistical tests are available for these purposes. While reporting the whole distribution may be a viable strategy where the number of response categories is relatively limited (e.g. example shown in Box 4.1), as the number of categories increase it becomes more difficult to reach overall judgements from purely descriptive data.

Box 4.1. **Reporting on the proportion of respondents by response category**

Statistics New Zealand publishes a number of measures of subjective well-being in the statistical releases for the biannual New Zealand General Social Survey. These include overall life satisfaction and satisfaction with particular aspects of life, namely financial satisfaction and a subjective assessment of health status. In all cases a five-point labelled Likert scale is used for responding to the questions. Although such a measure is sub-optimal in many respects, it lends itself well to being presented as a proportion of respondents by response category (Figure 4.1).

Figure 4.1. **Reporting the proportion of respondents selecting each response category**



Source: Statistics New Zealand, *New Zealand General Social Survey*.

One way to manage a large number of scale responses is to report on *the proportion of responses falling above or below a given threshold, or set of thresholds*. For example, responses can be reported as the percentage of the sample falling above or below a certain cut-off point, or banded into "high", "medium" and "low" categories (Box 4.2). Threshold descriptions of the data can be grasped quickly – providing an anchor for interpretation, and offering a way of

Box 4.2. **Output presentation examples – threshold-based measures**

The Gallup-Healthways Life Evaluation Index classifies respondents as "thriving", "struggling", or "suffering", according to how they rate their current and future lives (five years from now) on the Cantril Ladder scale with steps numbered from 0 to 10, where "0" represents the worst possible life and "10" represents the best possible life. "Thriving" respondents are those who evaluate their current state as a "7" or higher and their future state as "8" or higher, while "suffering" respondents provide a "4" or lower to both evaluations. All other respondents are classified as "struggling". Table 4.2 shows thriving struggling and suffering in the EU.

Table 4.2. **Gallup data on thriving, struggling and suffering in the EU (sorted by percentage suffering)**

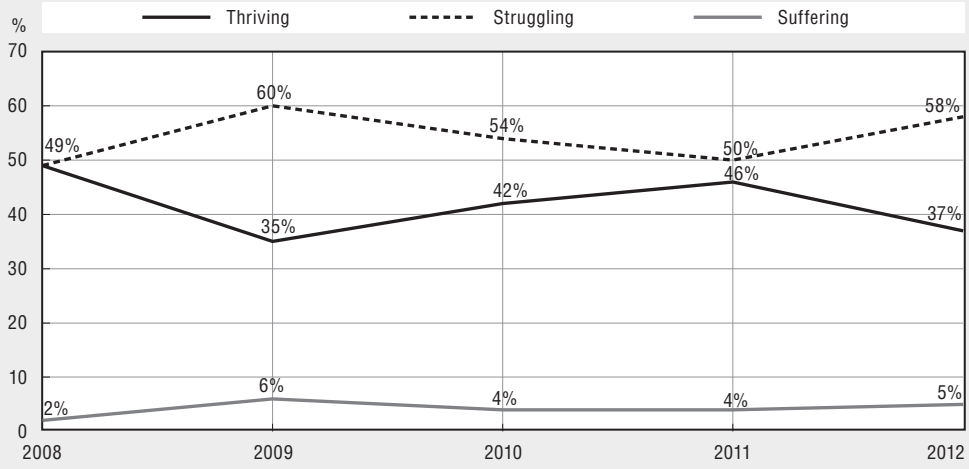
Column 1	% thriving	% struggling	% suffering	% thriving minus % suffering (pct. pts)
Bulgaria	5	50	45	-40
Romania	18	54	28	-10
Hungary	15	57	28	-13
Greece	16	60	25	-9
Latvia	16	61	23	-7
Portugal	14	65	22	-8
Estonia	24	60	17	7
Poland	23	60	17	6
Lithuania	23	57	16	7
Slovenia	32	53	14	18
Germany	42	52	6	36
Czech Republic	34	53	13	21
Slovak Republic	27	61	12	15
Malta	34	55	11	23
Spain	39	54	7	32
Cyprus	44	49	7	37
Italy	23	71	6	17
United Kingdom	52	44	6	46
Ireland	54	43	4	50
France	46	50	4	42
Austria	59	38	3	56
Finland	64	34	3	61
Denmark	74	24	2	72
Luxembourg	45	54	1	44
Netherlands	66	33	1	65

Note: Data collected between March and June 2011. Data unavailable for Sweden and Belgium at time of publishing.
Source: Gallup World web article by Anna Manchin, 14 December 2011, "More suffering than thriving in some EU countries", *www.gallup.com/poll/151544/Suffering-Thriving-Countries.aspx*.

Box 4.2. **Output presentation examples – threshold-based measures** *(cont.)*

Change in subjective well-being over time can also be presented relative to a given threshold (Figure 4.2).

Figure 4.2. **Share of the French population classified as "thriving", "struggling" and "suffering"**



Source: Gallup World web article by Anna Manchin, 4 May 2012, "French Adults' Life Ratings sink in 2012", *www.gallup.com/poll/154487/French-Adults-Life-Ratings-Sink-2012.aspx*.

communicating something about the distribution of the data with a single figure. The use of thresholds is also consistent with the ordinal nature of much subjective well-being data, as it requires no assumptions about the cardinality of scale responses.

The downsides of threshold measures include losing some of the richness of the data,[7] and the risk of encouraging a distorted emphasis on shifting people from just below to just above a threshold. This is a particular risk if only one threshold (e.g. "6 and above") is used, because it may be important for policy-makers in particular to understand what characterises communities at both high and low ends of the subjective well-being spectrum. Although thresholds have the potential to be more sensitive to change when carefully selected around the area of greatest movement on the scale, there is a considerable risk that a threshold positioned in the *wrong* part of the scale could mask important changes in the distribution of the data. For example, if the risk of clinically-significant mental health problems is greatest for individuals scoring 5 or less on a 0-10 life evaluation measure, setting a threshold around 7 could lead to a failure to identify changes that could have significant consequences for policy. In addition, reporting based on thresholds runs the risk of presenting two very similar distributions as quite different, or vice versa. For example, for some countries the distribution of subjective well-being is bi-modal, while for others there is a single mode. Depending on where a threshold is set, two such distributions might be presented as very different, or essentially the same. The central difficulty, therefore, lies in identifying *meaningful* threshold points that have real-world validity.

Thresholds can be set through examining the underlying distribution of the data and identifying obvious tipping points, but this data-driven approach limits both meaningful interpretation (what is the real-world meaning of a data cliff?) and comparability among

groups with different data distributions, whether within or between countries. A more systematic approach may be to adopt something similar to a "relative poverty line", whereby individuals falling, for example, below half of the median value on a scale are classified as faring badly. This capitalises on thresholds' ability to convey distributional characteristics, but has the downside of conveying relatively little about the average level, which is essential for both group and international comparisons.[8] It also remains an essentially arbitrary method for identifying a threshold. The final option would be to select an absolute scale value below which individuals demonstrate a variety of negative outcomes (and an upper bound associated with particularly positive outcomes), based on the available empirical evidence. This would at least give the threshold some real-world meaning.

Blanton and Jaccard (2006) make a strong case for linking psychological metrics to meaningful real-world events, highlighting the risk of assigning individuals to "high", "medium" and "low" categories without justifying or evidencing what these categories *mean* in practice. In particular, they note the conceptual and practical problems associated with the intuitively appealing practice of "norming", i.e. setting threshold values based on the proportion of the sample falling above or below that threshold. For example, in an obesity reduction programme, if an individual's weight loss result was described as "high" because relative to others in the group they lost more weight, the clinical significance of the finding remains obscured: it is possible that *everyone* in the group lost a clinically significant amount of weight, or *no-one* in the group lost a clinically significant amount. In both of these scenarios, what matters is not how the individual fares relative to the rest of the sample, but how their weight loss is likely to relate to other health outcomes. There is a clear analogy here with both relative poverty lines and international comparisons of subjective well-being: what would be categorised as "high" life satisfaction by normative standards in Denmark will be quite different to "high" life satisfaction according to normative standards in Togo – making these two categorisations impossible to compare. This emphasises the challenges associated with setting suitable thresholds and suggests *against* emphasising threshold-based measures too strongly in data releases. Given the wide range of potential uses of the data, a wide range of thresholds may be also relevant to policy-makers and others.[9]

Summary statistics of central tendency provide a useful way of presenting and comparing the level of subjective well-being in a single number. The most commonly-used measures of central tendency are the mean, the mode and the median. However, due to the limited number of scale categories (typically no more than 0-10), the median and modal values may lack sensitivity to changes in subjective well-being over time or to differences between groups. The mean is therefore generally more useful as a summary statistic of the level of subjective well-being.

Although the mean provides a good summary measure of the level of subjective well-being, it has shortcomings. First, the use of the mean requires treating the data from which it is calculated as cardinal. Although most subjective measures of well-being are assumed to be ordinal, rather than cardinal,[10] evidence suggests that treating them as if they were cardinal in subsequent correlation-based analysis does not lead to significant biases: the practice is indeed common in the analysis of subjective well-being data, and there appear to be few differences between the conclusions of research based on parametric and nonparametric analyses (Ferrer-i-Carbonell and Frijters, 2004; Frey and Stutzer, 2000; Diener and Tov, 2012). That said, Diener and Tov also note that when it comes to simpler analyses, such as comparisons of mean scores, ordinal scales that have been

adjusted for interval scaling using Item Response Theory can produce different results to unadjusted measures (p. 145). Second, the mean can be strongly affected by outliers and provides no information on the distribution of outcomes. Both of these issues therefore highlight the importance of complementing the mean with information on the distribution of data.

Distribution

It is also important to present information on the distribution of responses across the different response categories. If the primary way of presenting the data is by reporting the proportion of responses falling in each response category, the need for separate measures of distribution is less important. If, however, reporting is based on thresholds or summary statistics of central tendency, specific measures of distribution are important. The choice of distributional measure will depend partly on whether the data is treated as ordinal or cardinal.

When cardinality is assumed, it is possible to use summary statistics of distribution such as the Gini coefficient. Both the Gini coefficient and the standard deviation are based on calculations that are unlikely to hold much meaning for the general public, and may therefore be less effective as a tool for public communication. The Gini in particular also perhaps has less *intuitive meaning* for subjective well-being than it does for its more traditional applications to income and wealth.[11] This means that other measures of dispersion, such as the interquartile range (i.e. the difference between individuals at the 25th percentile and individuals at the 75th percentile of the distribution), or the point difference between the 90th and the 10th percentile (Box 4.3), may be preferred in simple data releases. Where space allows, graphical illustrations of distribution are likely to be the most intuitive way to represent distributions for non-specialist audiences, although such graphs can be difficult to compare in the absence of accompanying summary statistics.

Aggregation of multi-item measures

Where a survey includes more than one question about subjective well-being, a key reporting decision for data producers will be whether to report responses to each question separately, or alternatively to aggregate some questions into broader multi-item measures. Single-item life evaluation questions are most often reported as stand-alone headline measures.[12] However, in addition to the single-item life evaluation primary indicator, the suite of question modules proposed in Chapter 3 also includes several multi-item measures intended to capture evaluative, affective (or hedonic), eudaimonic and domain-specific aspects of subjective well-being.

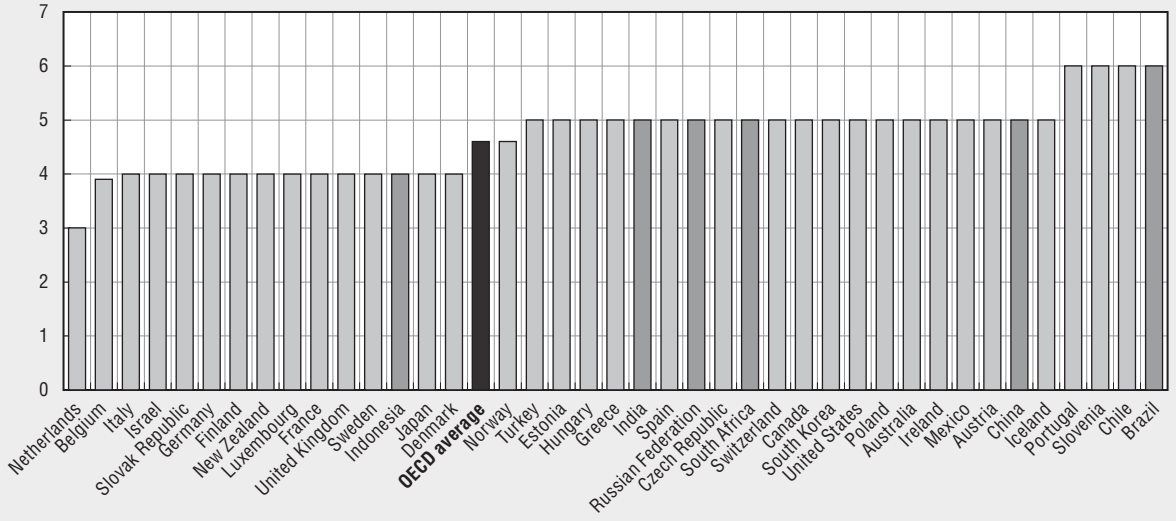
Although there may be value in looking at responses to individual questions or scale items in more detailed analyses, it is desirable to summarise longer multi-item measures, particularly for the purposes of reporting outcomes to the general public. Furthermore, summing responses across multiple items should generally produce more reliable estimates of subjective phenomena, reducing some of the impact of random measurement error on mean scores – such as may result from problems with question wording, comprehension and interpretation or bias associated with a single item. However, whilst summing responses across different life evaluation questions should pose relatively few problems, affect and eudaimonia are by nature more multidimensional constructs, and thus there is a greater risk of information loss when data are aggregated.

Box 4.3. **Distribution of subjective well-being among OECD and emerging countries (OECD, 2011a)**

In *How's Life?*, the OECD used the gap between the 10th and 90th percentiles as a measure of distribution (the "90/10 gap"). Conceptually similar to the interquartile range, the 90/10 gap was used because the clustered nature of life satisfaction responses meant that the interquartile range provided little to distinguish between countries.

Figure 4.3. **Inequality in life satisfaction in OECD and emerging economies, 2010**

Point difference between the 90th percentile and the 10th percentile



Source: Gallup World Poll data, reported in *How's Life?* (OECD, 2011a).

Options for aggregation, specific to each scale, include:

- *Positive and negative affect:* Where several items are used to examine experienced affect, most scales are designed such that one can calculate positive and negative affect subtotals for each respondent, summarising across items of similar valence. For example, in the core affect measure proposed in Module A of Chapter 3, positive affect is calculated as the average score (excluding missing values) for questions on "enjoyment" and "calm", and negative affect is calculated as the average score for questions on "worry" and "sadness". As with any summary measure, this risks some degree of data loss, particularly where affect dimensions can be factored into one or more sub-dimensions – for example, the high-arousal/low-arousal dimensions identified in the Circumplex model of mood (Russell, 1980; Russell, Lewicka and Niit, 1989; Larsen and Fredrickson, 1999). However, for the purposes of high-level monitoring of affect, examining summary measures will be more feasible than looking at each affect item individually, and the increased reliability of multi-item scales will be advantageous.
- *Affect balance:* Positive and negative affect measures can be further summarised into a single "affect balance" score for each respondent by subtracting the mean average negative affect score from the mean average positive affect score. This can then in turn be reported as either a mean score (positive minus negative affect) or as a proportion of the population with net positive affect overall.

- Where information is available on the frequency of positive and negative affect experiences throughout the day, such as that provided by time-use studies, it is also possible to calculate the proportion of time that people spend in a state where negative affect dominates over positive affect. This is described as the "U-index" (Kahneman and Krueger, 2006), and again this can also be reported at the aggregate population level. Time-use data also enable the mean affect balance associated with different activities to be described (Table 4.3).

Table 4.3. **Mean net affect balance by activity, from Kahneman et al. (2004)**

Activity	Percentage of sample	Time spent (hours)	Net affect[1]
Intimate relations	11	0.21	4.74
Socialising after work	49	1.15	4.12
Dinner	65	0.78	3.96
Relaxing	77	2.16	3.91
Lunch	57	0.52	3.91
Exercising	16	0.22	3.82
Praying	23	0.45	3.76
Socialising at work	41	1.12	3.75
Watching TV	75	2.18	3.62
Phone at home	43	0.93	3.49
Napping	43	0.89	3.27
Cooking	62	1.14	3.24
Shopping	30	0.41	3.21
Computer at home	23	0.46	3.14
Housework	49	1.11	2.96
Childcare	36	1.09	2.95
Evening commute	62	0.62	2.78
Working	100	6.88	2.65
Morning commute	61	0.43	2.03

1. Net affect is the average of three positive adjectives (enjoyment, warm, happy) less the average of five negative adjectives (frustrated, depressed, angry, hassled, criticised). All the adjectives are reported on a 0-6 scale, ranging from "not at all" to "very much". The "time spent" column is not conditional on engaging in the activity. The sample consists of 909 employed women in Texas.

Source: Kahneman, Krueger, Schkade, Schwartz and Stone (2004), Figure 2, p. 432.

Both affect balance and the U-index are similar to threshold-based measures, but ones that have both clear meaning and the considerable advantage of reducing affect data to a single variable. However, there is some risk of data loss in adopting these aggregation approaches, particularly when exploring group differences. For example, the ONS subjective well-being data release (ONS, 2012) found that, for most age groups, on average women reported slightly higher happiness yesterday than men, but they also reported higher anxiety yesterday. If aggregated as an affect balance measure, these differences may not be detectable.

Ultimately, the judgement of the most appropriate measure should be driven by the primary data use. For overall monitoring, the benefits of reporting affect balance are likely to outweigh the drawbacks – but when attempting to understand, for example, the links between affect and health outcomes, it may be more important to examine dimensions of affect separately (Cohen and Pressman, 2006).

Eudaimonia: Most of the literature regards eudaimonia as a multidimensional construct (e.g. Huppert and So, 2011; Ryff, 1989; Ryan and Deci, 2001), and therefore summarising across all items on a multi-item scale again risks some data loss. For detailed analysis, it may be important to examine each sub-component of eudaimonia separately, at least initially. Nonetheless, for the purposes of monitoring well-being, if positive correlations are found between each of the sub-dimensions, it may be appropriate to sum across items.[13] The first option is to take the mean average value of all responses, omitting missing values. Alternatively, a threshold-based approach has been proposed by Huppert and So (2011; see Box 4.3), which categorises respondents according to whether they meet the criteria for "flourishing". The "flourishing" construct may offer a powerful communicative device. However, partly because it is based on groups of items with different numbers and different response categories, Huppert and So's operational definition of "flourishing" ends up being quite complex (with different thresholds being applied to differentially distributed data, and different items grouped according to various subscales assumed to be present). As noted earlier, the present difficulty with threshold-based measures is that there is little consensus on where the meaningful cut-off points lie. Further research is therefore needed before this approach can be regarded as preferable to reporting mean average scores.

Domain satisfaction: Questions about satisfaction with individual domains of life can be meaningful as stand-alone measures, and may be particularly useful for policy-makers seeking specific information on the effects of a particular policy intervention. However, some sets of domain-specific questions have been designed with a view to creating a composite measure of life evaluation overall, by summing responses across each of the domains (e.g. the Australian *Personal Wellbeing Index* – International Wellbeing Group, 2006, in Module E, Chapter 3). This overall approach requires making strong assumptions about the weights to apply to each life domain (as well as the universality with which those weights apply across the population) along with some judgements about which domains of life are relevant to subjective well-being overall. In the case of the *Personal Wellbeing Index*, domains have been selected as the most parsimonious list for capturing "satisfaction with life as a whole", and equal weights are adopted for each domain, in recognition of the fact that empirically-derived weights may not generalise across data sets. These assumptions notwithstanding, composite measures of domain satisfaction may offer a more rounded and potentially more reliable picture of life "as a whole", as respondents are encouraged to consider a variety of different aspects of life when forming their answers.

Aggregating several subjective well-being indicators into an overall index

Although the various subcomponents of subjective well-being (e.g. life evaluation, eudaimonia and affect) will convey most information when measured and reported separately, there may be demand for aggregating these into a single over-arching index of subjective well-being, particularly for the purposes of high-level communication and monitoring[14] (see Stiglitz, Sen and Fitoussi, 2009, for a detailed discussion of aggregation issues in relation to well-being indicators). Where there is pressure to report just one overall headline measure, selecting only one element of subjective well-being (such as life evaluations) may neglect other important components, making aggregation across life evaluations, eudaimonia and affect an attractive prospect to those who wish to see all three components of subjective well-being reflected in headline measures.

The communication advantages in reducing different measures of subjective well-being to one number must, however, be set against a number of methodological objections. Most fundamentally, the different aspects of subjective well-being (life evaluation, affect, eudaimonia) represent distinct constructs, and it is not clear that it is possible to provide a coherent account of what an aggregate index of overall subjective well-being actually represents. Similarly, there is no clear basis for determining the relative weights to assign to different dimensions or sub-dimensions of subjective well-being. This problem is analogous to those encountered when trying to develop composite measures for other well-being outcomes, such as health or skills. Until further consideration has been given to how composites could be created, the most sensible approach may be for data producers to provide disaggregated measures – enabling users to experiment and create their own composite indices as necessary. In the meantime, where single headline figures are to be reported, life evaluations are likely to remain the focus, because they are currently the most established of the three measures in terms of their use to complement existing measures of well-being (see Chapter 1).

Reporting change over time and differences between groups

National levels of subjective well-being are difficult to interpret when examined in isolation. External reference points are essential in order to understand whether a mean life satisfaction score of 7.2 is "good", or not. In order to interpret current observations of subjective well-being, two broad comparisons are likely to be of interest to data users: 1) comparisons between current and previous levels of subjective well-being; and 2) comparisons between different countries, particularly those regarded as peers in terms of their overall levels of development.

A third type of comparison involves examining group differences *within* a country. Identifying groups of individuals who report lower or higher subjective well-being, or whose well-being is changing at a faster or slower rate over time, is an essential use of national statistics. Defining reference groups for such comparisons is also important – and by providing information about the level of subjective well-being across the whole population, national statistics provide a baseline against which population sub-groups can be compared. Further breakdowns in national statistics (such as by age, gender, education, region, occupation, socio-economic and employment status, health status, etc.) can also enhance their usefulness. Understanding what characterises communities at both high and low ends of the subjective well-being spectrum will be important for policy users seeking both to reduce extreme suffering and to better understand how high levels of subjective well-being can be achieved.

Examining whether gaps in subjective well-being between groups within society are growing or shrinking is also important. Central and local governments, the wider public sector, researchers and voluntary organisations may be particularly interested in inequalities in subjective well-being in order to assist the identification of vulnerable groups who may benefit from specific interventions.

Comparisons over time and between groups, both within and across countries, can also signal where to look in terms of the potential *drivers* of subjective well-being. For example, if regional differences in subjective well-being are identified, looking at other variables which differ across regions can have implications for better understanding what matters for subjective well-being. This will be of interest to government and researchers, but also to members of the public and the organisations that they work for.

Methods for reporting change over time and differences between groups

There are also several ways in which comparisons over time and between groups can be reported. The first step involves basic descriptive statistics. These include tracking mean changes in time series, calculating changes in the mean score between time points, examining absolute or percentage differences between groups, and looking at group differences over time or relative to a given threshold (see Figures 4.2 and 4.3, Box 4.2). Changes in the overall *distribution* of subjective well-being over time are also of interest, as they can indicate whether society as a whole is becoming more or less equal in terms of people's experiences of subjective well-being. Finally, differences in the rate and direction of change both between groups within societies and between countries more broadly may also be important.

Data users will find summary statistics easier to understand if there is some degree of consistency between the methods used for reporting current levels of subjective well-being and those used for reporting change over time or comparisons between groups. Thus, if current levels are described using the mean, ideally change over time should also be reported on this basis. Once again, threshold-based estimates offer both advantages and disadvantages. Ease of communication and sensitivity to changes around the threshold level come at the cost of failing to detect changes or differences elsewhere in the scale.[15] Although the overall information loss can in theory be managed through careful selection of the threshold value (and potentially through multiple thresholds), it is not obvious where that threshold should be drawn. Selecting cut-offs according to the distribution of the data could result in setting different thresholds for different population groups and/or different countries, making comparisons impossible.

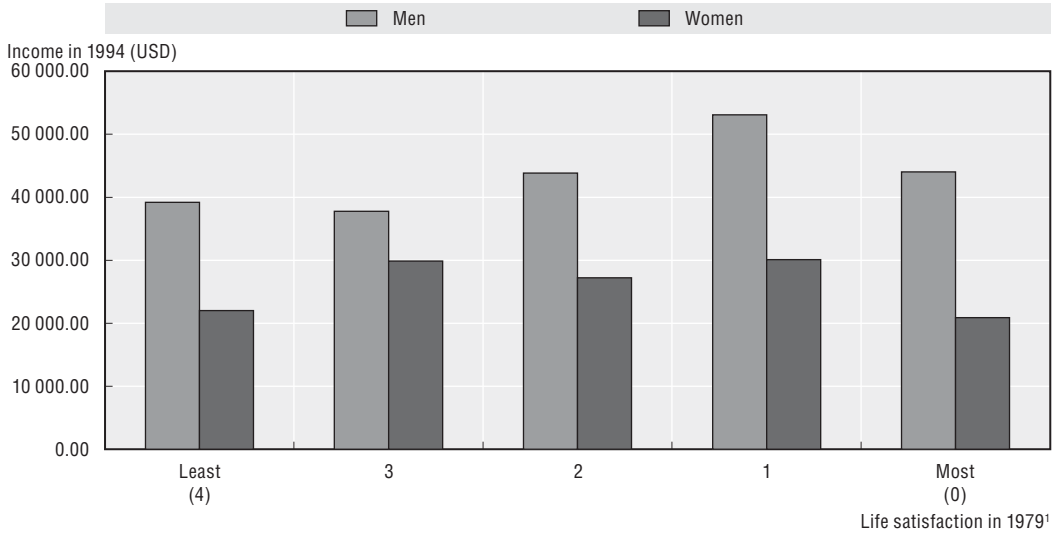
A number of factors can make comparisons of basic descriptive statistics challenging to interpret: for example, differences in sample sizes, or the variability of the data, can make simple comparisons between summary statistics misleading. Thus, both the sample size and standard errors (i.e. the standard deviation of the sampling distribution of a statistic) should be considered when comparing two or more different observations – whether over time or between groups. Robust estimates of standard errors require large and representative samples: when sample sizes are small, standard errors can be larger and the risks of false inferences greater.

One approach is to ensure that whenever group means are reported, both the group sample size and the standard deviations are reported alongside. To assist in the interpretation of standard errors, it may be preferable to display these graphically, for example through the use of box-plots or of error bars added to charts comparing mean levels (Figure 4.4), or simply by providing bar charts to show any differences in the distribution of the data in each group. Finally, statistical inference testing offers a way to examine the likelihood that the observed difference between two values would occur by chance – taking both sample size and standard errors into account.

Analysing and interpreting descriptive subjective well-being data

Almost all analysis associated with monitoring progress will be concerned with examining differences between observations. Whilst statistical analyses can provide a sense of the *statistical* significance of an observed difference, they cannot indicate the *practical* significance of a finding – both in terms of its overall size (*is this difference big enough*

Figure 4.4. **Australian longitudinal study: Life satisfaction and income (Oishi et al., 2007)**



Note: Data drawn from the Australian Longitudinal Study. Data are from men and women surveyed in 1979 (for life satisfaction) and again in 1994 (for income). Along the x-axis, respondents (men and women) are grouped according to the level of life satisfaction reported on a 5-point scale, from most satisfied (0) to least satisfied (4). The y-axis shows mean average reported income in 1994 for each of the respondent groups. Error bars represent standard error.
1. 0 = most satisfied; 4 = least satisfied.
Source: Oishi et al., *Perspectives on Psychological Science, 2007*, No. 2, pp. 346-360.

to matter?) and the possible source of the difference observed (*how can we know if the difference is genuine?*). The section that follows examines three central issues when seeking to interpret patterns of subjective well-being over time or among groups:

- What size of difference can we expect to see?
- What alternative explanations should be considered for the observed differences?
- What is the role of culture in international comparisons, and can data be "corrected" for "cultural bias"?

For subjective well-being to be a useful complement to other measures of well-being, it needs to reflect changes in the things that matter to people. While there is clear evidence that life circumstances have a significant impact on subjective well-being levels, the average measures for countries generally appear to change very slowly over time, and sometimes only by small amounts in response to quite substantial events. In contrast, differences between countries can sometimes appear large relative to what is known about how those countries differ solely on economic measures of well-being.

Even when making very simple comparisons over time, among groups, or among countries, it is important to consider the possible drivers and alternative explanations for observed differences over time or among groups. One particular source of concern is the potential for cultural "bias" to influence cross-country comparisons.

What size differences can be expected?

Despite the wide variety of factors that can limit the size of differences observed in subjective well-being data (considered below), evidence clearly shows that measures can and do change in response to life circumstances (Veenhoven, 1994; Lucas, 2007a, 2007b; Lucas, Clark, Georgellis and Diener, 2003; Diener, Lucas and Napa Scollon, 2006). The expected size of differences in subjective well-being measures, however, varies depending on the context of the analysis. What might be considered a "medium-sized" difference between two population sub-groups within a country would constitute a "large" change in the average level within a country over time, but only a "small" difference between countries.

In addition to considering differences in the mean levels of subjective well-being, it can be valuable to consider the standard errors associated with estimates for different groups, for different observations over time and for different countries. This is currently overlooked in a number of reports, but is important both to understand the likely robustness of mean differences and to better indicate the distribution of the data. The inequality of subjective well-being within groups and across society can be an important indicator, and evidence also suggests that individuals' subjective well-being can vary considerably in response to certain life events, such as disability (Diener, Lucas and Napa Scollon, 2006; Schulz and Decker, 1985). This makes the standard errors of mean estimates particularly relevant.

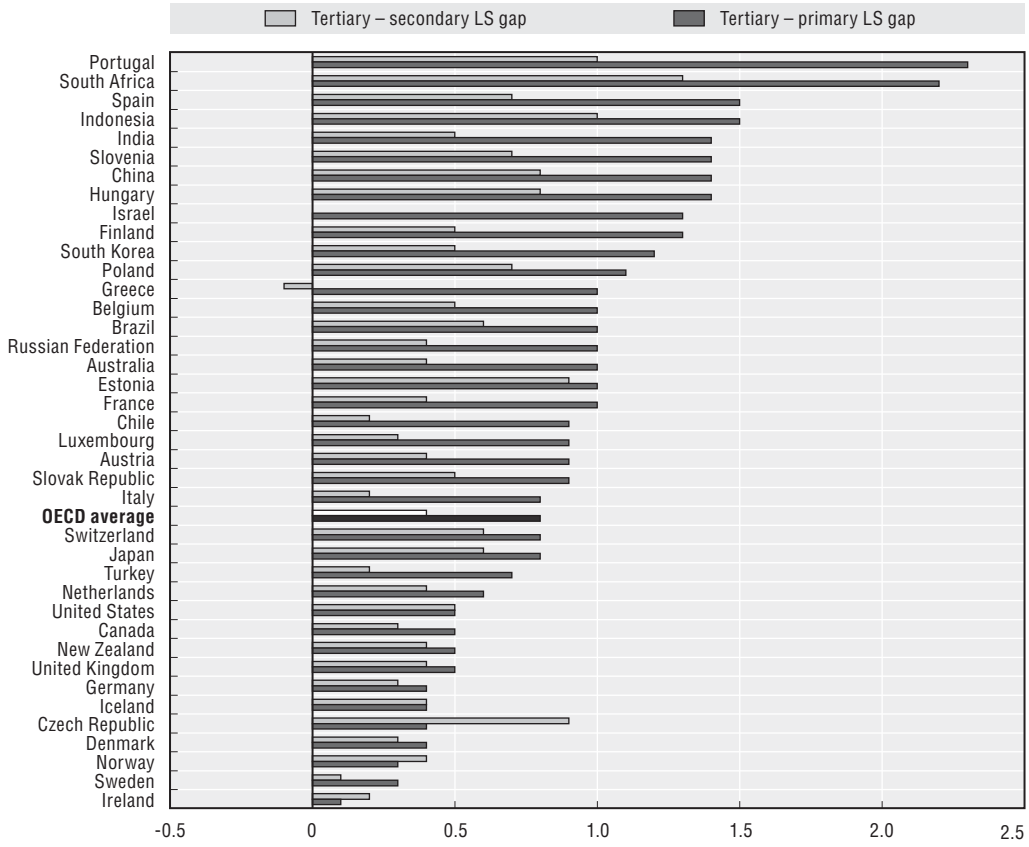
Differences among groups. Within affluent countries, simple mean differences in the range of around 0.5 to 2 scale points on a 0-10 scale (a 5-20% difference) have been detected among different population sub-groups on life evaluation, eudaimonia and affect measures. For example, analysis of experimental subjective well-being data collected from a nationally-representative sample of 80 000 United Kingdom adults in 2011 (ONS, 2012) found differences between employed and unemployed respondents of around 1 scale point in response to life evaluation and eudaimonia questions, and around half a scale point in response to "happy yesterday" and "anxious yesterday" questions (all measured on a 0-10 scale). Similar magnitude differences were observed between those married or in a civil partnership, and those who were divorced or separated – although mean affect differences were closer to 1 scale point between these groups.

Health is another important component of well-being overall, and reductions in subjective well-being have also been observed for groups experiencing health problems. The ONS (2012) reported mean life satisfaction, eudaimonia, and "happy yesterday" responses between 1.7 and 2.0 scale points lower, and "anxious yesterday" responses 1.7 scale points higher, among those out of work due to long-term sickness, in comparison to total population means (all measured on a 0-10 scale). Lucas (2007a) reports data from two very large-scale nationally-representative panel studies several years *before and after* the onset of a disability. In this work, disability was associated with moderate to large drops in happiness (with effect sizes ranging from 0.40 to 1.27 standard deviations), and little adaptation over time.

The OECD (2011a) also reports gender differences in life evaluations and affect balance. While in the United States, Japan, Finland and China women report higher average levels of both life evaluation and affect balance than men (with the ratio of men's scores to women's in the range of 0.90-0.99), in Eastern and Southern Europe, Latin America and the Russian Federation men are more likely to report positive affect balance and higher life evaluations. This difference is most marked in the cases of Hungary, Slovenia and Italy, where the ratio of men's scores to women's is 1.05 or above for life

evaluations, and 1.15 or above in the case of positive affect balance. Looking at education gaps, low levels of education are associated with lower levels of life evaluations overall, but in Portugal, Spain, Slovenia, Hungary and many of the non-OECD countries for whom data is provided there are particularly large differences in life evaluations between people with and without tertiary education (Figure 4.5).

Figure 4.5. **Gap in life satisfaction by level of education for OECD and selected countries, 2010**



Note: The gap is defined as the difference between the mean life satisfaction of people with tertiary attainment and the mean life satisfaction of people with primary (secondary) education.
Source: OECD's calculations on Cantril Ladder (0-10 scale) life evaluations Gallup World Poll, reported in *How's Life?* (2011a).

Between-country comparisons. In analyses of subjective well-being among different countries, life evaluations appear to vary by up to 5 scale points (on a 0-10 scale) globally, and affect balance by up to 0.5 on a 0-1 scale, although it is important to note that the available evidence on global differences is currently based on small and in some cases unrepresentative samples. The *World Happiness Report* (Helliwell, Layard and Sachs, 2012, Figure 2.3) describes mean average life evaluations averaging 7.6 (on a 0-10 scale) for the top four countries (Denmark, Finland, Norway and the Netherlands), whereas mean reported levels fall well below 4 in the bottom four countries (Togo, Benin, Central African Republic and Sierra Leone). Differences of up to 0.5 are reported (on a 0-1 scale) for positive affect balance, with respondents in Iceland, Laos and Ireland reporting an average positive affect balance of around 0.7, and those in the Palestinian Territories, Armenia and Iraq reporting an average below 0.2.

Even among relatively affluent societies, differences in levels of subjective well-being are non-trivial. Among OECD countries, the mean average life evaluations (on a 0-10 scale) reported in 2010 ranged from over 7.5 in the cases of Denmark, Canada and Norway to between 5.0 and 6.0 in the cases of Portugal, Estonia, Turkey, Greece and Poland (OECD, 2011a). In terms of the proportions of the population experiencing a positive affect balance,[16] Denmark, Iceland, Japan, the Netherlands, Norway and Sweden each score more than 85%, whilst Turkey, Italy, Israel, Portugal and Greece score around 70%.

Both the OECD (2011a) and Helliwell, Layard and Sachs (2012) also report substantial differences between countries in the distribution of subjective well-being. The standard deviations reported for life evaluations on the 0-10 Cantril Ladder reported by Helliwell et al. were around or below 1.5 for the Netherlands, Denmark, Finland and Belgium, indicating quite consistently high levels of subjective well-being. However, similar standard deviations were also observed for Namibia, Tajikistan, Senegal, Madagascar, Côte d'Ivoire, Niger, Cambodia, Burkina Faso, Chad, Comoros, Burundi and the Central African Republic, which among these countries indicates quite consistently low levels of subjective well-being. While the mean life evaluation scores for Puerto Rico and Colombia may seem quite high, considering their levels of economic development, variability among these scores is also high, with standard deviations of around 2.6 and 2.5 respectively. Marked variations in life evaluations are also evident for countries such as Honduras, Pakistan, Nicaragua, Lebanon and the Dominican Republic. These relatively high standard deviations invite more research into their sources, permanence and consequences.

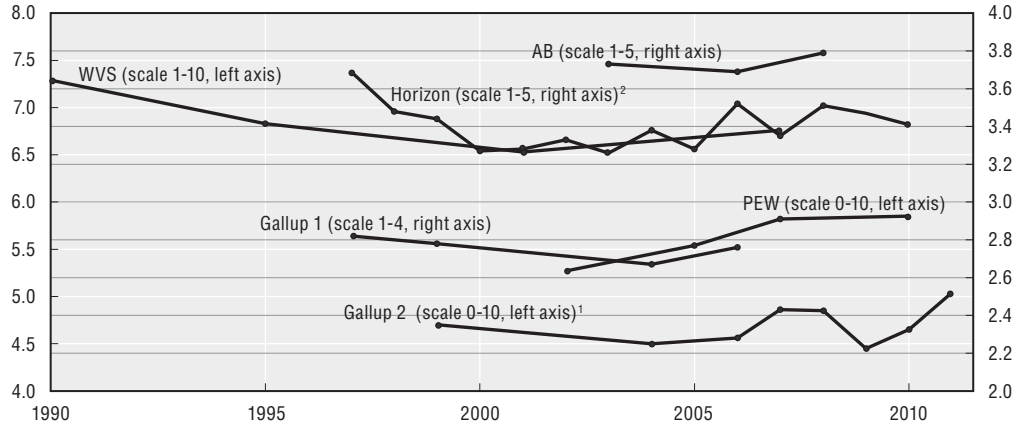
Changes over time. When looking at changes in average country scores over time, the level of change to expect will be highly dependent on the extent of social, political and economic change taking place – and the combined effect of several different changes needs to be considered collectively when interpreting the overall pattern of subjective well-being. Evidence suggests that at the national aggregate level, a *long-term*[17] mean average shift of 0.3 or 0.5 scale points on a 0-10 life evaluation scale may represent a very sizeable change, occurring only in response to major societal shifts. If one bears in mind that little more than 4 scale points separate the top and bottom national life evaluations among the 150 countries in the Gallup World Poll, a shift of 0.4 points would change a country's international ranking by ten to twenty places.

Several authors have examined the question of whether increases in income over time are associated with increases in subjective well-being, and particularly increases in life evaluations (e.g. Easterlin, 1974, 1995, 2005; Hagerty and Veenhoven, 2003; Sacks, Stevenson and Wolfers, 2010). Frijters, Haisken-DeNew and Shields (2004) looked at the effect of the large increase in real household income in East Germany on life satisfaction following the fall of the Berlin Wall in 1989. In the 10-year period between 1991 and 2001, the authors estimated that around 35-40% of the observed increase in average life satisfaction was attributable to the large (over 40%) increase in real household incomes during this time period, with a one-unit increase in log income corresponding to around a 0.5 unit increase in life satisfaction for both men and women.

However, it is important to consider other social and political changes that may be co-occurring when examining the effects of increasing income – such as the potential impact of growing inequalities in society. For example, Easterlin et al. (2012) chart the changing levels of overall life evaluations in China between 1990 and 2010, a period during which China's GDP per capita increased fourfold. According to World Values Survey data

analysed by these authors, there was a *decline* in life satisfaction of 0.53 scale points on a 1-10 scale from 1990 to 2007, although an upturn begins to emerge from early 2000 onwards. Although the early World Values Survey data points in this analysis may be biased upwards, the same general "U-shape" with a turning point between 2000 and 2005 is visible in other data (Figure 4.6).

Figure 4.6. **Easterlin et al. 2012: China's life satisfaction, estimated from six time series data sets, 1990-2010**



1. Scale of 1-10 for 1999 and 2004 data.
2. For 1997, 1998, 1999 and 2001: Scale of 1-4 and mean computed from 5, 4, 2, 1 coding.

Source: Easterlin, Morgan, Switek and Wang (2012), p. 2.

In conclusion, what constitutes a "big" difference in subjective well-being partly depends on the nature of the difference under consideration. Although results from individual studies need to be interpreted with caution, current evidence suggests that within countries group differences in mean scores of around 10% may be considered large, whereas between countries much greater differences can be observed because the cumulative impacts of different life circumstances can stack on top of one another. Looking at change within countries over time, it is important to consider a wide variety of potential drivers when interpreting the results – because a smaller change than might be expected, or an unexpected direction of change, may be due to the combined and possibly interacting effects of a number of variables. It is also very valuable to examine differences in the *distribution* of subjective well-being data, both over time and between groups. This important consideration is often neglected, but deserves closer attention.

What influences effect sizes?

Several factors potentially limit the size of the difference between groups or change over time that one can expect to see in subjective well-being data. These include the boundedness of the scale and the overall distribution of responses; the sample size in each group and the proportion of the sample affected by a given societal change; changes or differences in other important determinants of well-being; the influence of frame-of-reference effects; the possibility that subjective well-being may be both a cause and an effect of group differences and societal changes; and the time frame over which differences or changes are examined. When interpreting the magnitude of changes in subjective well-being over time, or differences among population sub-groups, it is important to consider how these factors are influencing estimates, particularly when these deviate from the expected pattern.

Bounded scales, response categories and the distribution of responses. One practical issue that potentially limits movements over time is that subjective well-being data are collected using bounded scales with a limited number of response categories. Unlike some indicators of well-being where the scale is unbounded (e.g. income; life expectancy), the average subjective well-being response can never move beyond the top response category. Furthermore, where the choice of response categories is very limited (such as the three and four response categories in the happiness questions often used to investigate the Easterlin paradox), a quite massive shift in life circumstances may be required to move individuals up or down by even just one scale point. Bradburn, Sudman and Wansink (2004) suggest that, on a three-item scale with extremes at either end (for example, "best ever" "worst ever" and "somewhere in between"), most people will tend to select the middle category. This scale sensitivity issue is a key part of the rationale behind preferring longer response scales with a greater number of categories (discussed in depth in Chapter 2). Among highly-developed countries, life evaluations and affect data also tend to be skewed so that the large majority of responses are in the upper range. Whilst at present even the highest-scoring countries still report mean scores several points below the scale ceiling on 0-10 measures, it is rare for respondents to declare their lives or affective states as being absolutely perfect. This implies that "quick wins" in terms of improving overall subjective well-being may be relatively few and far between, and that those high-ranking countries seeking to further improve overall well-being may best focus their energies on the lower tail of the distribution. However, "quick losses" in mean scores of subjective well-being also still remain possible, as highlighted in the experience of Egypt and Tunisia in the years prior to the Arab Spring.

Despite skewed distributions in life evaluation and affect in affluent countries, in emerging and developing economies there is considerable scope for subjective well-being to improve (as documented in the recently published *World Happiness Report*, Helliwell, Layard and Sachs, 2012[18]). In the case of eudaimonia, Huppert and So (2011) examined data from 23 European countries and found that even in the best-performing country (Denmark) only 41% of the population were considered to be "flourishing". The next-best performer was Switzerland at 31%, whereas in Portugal fewer than 10% of the population met the criteria. Thus, whilst the nature of the subjective well-being response scales place theoretical limits on how high subjective well-being can ultimately go, the vast majority of countries do not currently appear to be anywhere near those limits.

Proportion of the sample affected. When interpreting national data, it is also important to consider that life circumstances or conditions that only affect a small percentage of the overall sample (or groups being compared) may have large effects at the individual level whilst having a relatively small effect on the aggregate level. For example, unemployment is known to have a powerful negative impact on the life evaluations of those individuals affected, and this is evident when comparing average scores for the unemployed and the employed. However, even relatively large increases in the unemployment rate (for example, from 5% to 10%) may lead to only small decreases in mean life evaluations for the country as a whole (e.g. Deaton, 2012), in part because they only affect a small proportion of the population. When combined with the high signal-to-noise ratio of subjective well-being indicators, this means that large samples are often needed to detect meaningful long-term shifts in subjective well-being over time at the national level.

The wide variety of subjective well-being determinants. The very large number of determinants of subjective well-being also means that changes over time in any one variable, or in any one difference between two groups of individuals, may have only a small impact on mean scores. The range of variables showing significant associations with subjective well-being includes health, income and material wealth, employment status, migrant status, education, marital status, social relationships, trust in others, volunteering, confidence in institutions, governance, freedom, air quality, personal safety and crime – to name just a few (Boarini et al., 2012). Thus, when examining either changes in country-level mean scores over time, or mean score differences between groups, it is important to consider *other variables* that may also be changing over time or differing between those groups, which could serve to reduce or obscure the effects of the variable in question.

Frame-of-reference effects and adaptation. A person's responses to questions about subjective well-being will be informed by the limits of his or her own experience. "Frame-of-reference effects" refer to differences in the way respondents formulate their answers to survey questions, based on their own life experiences as well as their knowledge about the experiences of others, including both those they consider as within their "comparison group" and those outside it (Sen, 2002; Ubel et al., 2005; Beegle, Himelein and Ravallion, 2012). This knowledge and experience sets the frame of reference, relative to which a respondent's own current circumstances and feelings are felt and evaluated.

Frames of reference produce real differences in how people genuinely feel, rather than simply differences in how people report those feelings. Thus, frame-of-reference effects do not bring into question the validity of subjective well-being measures *as measures of subjective constructs*, but rather they are concerned with the relationship between objective and subjective experiences. Framing effects matter when using subjective well-being as a complement to other measures of well-being, because they concern the extent to which subjective well-being is a *relative* construct, rather than something reflecting *absolute* achievements in society. However, the available evidence suggests that, while framing effects may influence the *size* of group and country differences observed in subjective well-being data, they are not sufficiently large to prevent the impact of life circumstances from being detected (e.g. Boarini et al., 2012; Fleche, Smith and Sorsa, 2011; Helliwell and Barrington-Leigh, 2010).

Adaptive psychological processes can also either restore or partially repair subjective well-being, and particularly affective experiences, in the face of some types of adversity (e.g. Cummins et al., 2003; Diener, Lucas and Napa Scollon, 2006; Clark et al., 2008; Riis et al., 2005). Adaptation to positive life events such as marriage or winning the lottery has also been observed (Clark et al., 2008; Brickman, Coates and Janoff-Bulman, 1978).

The possibility of shifting reference frames and psychological adaptation again mean that differences over time, between groups and between countries might be smaller than one might expect based on objective changes or differences in life circumstances. However, there is strong evidence that adaptation does not occur (or is incomplete) for a range of policy-relevant life circumstances, such as chronic pain from arthritis or caring for a severely-disabled family member (Cummins et al., 2003), disability (Oswald and Powdthavee, 2008a; Lucas 2007a; Brickman, Coates and Janoff-Bulman, 1978) and unemployment (Lucas et al., 2003). Focusing on instances of incomplete adaptation could help policy-makers and public service providers to focus on areas where intervention may be most valuable.

Reverse and two-way causality and time frames for analyses. The possibility of reverse or two-way causality[19] among subjective well-being and its determinants may also limit the size of difference likely to be observed in subjective well-being data. For example, at the cross-sectional level, there is evidence that being married is associated with higher levels of both life satisfaction and positive affect, and with lower levels of negative affect/anxiety (e.g. Boarini et al., 2012; ONS, 2012). However, there is also some evidence that happier people are more likely to get married (Luhmann et al., 2012), and that after the initial boost in subjective well-being observed around the time of marriage, subjective well-being reduces back to its pre-marriage levels in the years after the event (Clark et al., 2008). Thus, an increase in the proportion of the population getting married in year *n* may not necessarily produce a large increase in the average levels of subjective well-being reported in year *n* + 5. Conversely, an increase in average levels of subjective well-being may actually *precede* an increase in marriage rates.

Both reverse/two-way causality and adaptation raise the issue of the appropriate *time frame* to consider when examining changes in subjective well-being. Whilst some determinants of subjective well-being might be expected to have an immediate impact on feelings of well-being (such as the sudden onset of disability or the death of a family member), others may take longer to unfold because their effects are indirect (for example, the influence of education on subjective well-being, or of having sufficient income to enable investment in healthcare insurance or a pension for retirement). Thus, as with any attempt to evaluate impact, for sizeable differences to be detected the correct time-frame for analysis needs to be adopted, based on what is known about the variables in question and the causal pathways through which they take effect.

Although longer time-frames might be required to detect significant changes in subjective well-being data, it is also true that these measures can be relatively bumpy over short time periods. Deaton (2012), for example, raises the possibility that long-term trends in country-level subjective well-being risk being swamped by "cognitive bubbles", i.e. by the temporary impact of short-term reactivity to national events that affect everyone (such as public holidays or major news events). If time-series data on subjective well-being are examined over only short time periods, these bubbles can potentially drown out the more meaningful changes associated with important societal shifts in well-being (such as rising unemployment rates), particularly if these affect only a small proportion of the population during the time-frame examined. On the one hand, short-term measures may act as a useful barometer for public mood – and the short-term worry and stress that accompanied the immediate impacts of the 2008 financial crisis "is surely real enough, and worth measuring and taking into account in policy" (Deaton, 2008, p.23). On the other hand, it is important to view short-term fluctuations in subjective well-being in the context of much broader long-term trends in order to capture wider changes in what most people might regard as societal progress.

What alternative explanations should be considered for observed differences in subjective well-being?

Because subjective well-being is affected by so many different life circumstances, several factors need to be taken into account when interpreting the magnitude of a difference between groups or a change in subjective well-being data over time. A number of background characteristics, such as age, gender and marital status, can influence mean

levels of subjective well-being. When making group comparisons (for example, between the employed and unemployed, or between different regions within a country), it is important to consider the impact of differences in these background characteristics.

Regarding age differences, evidence suggests that among affluent OECD and especially English-speaking countries, there is a U-shaped relationship between age and life satisfaction (with average levels lowest between the ages of around 35 and 55), and this persists even after controlling for other age-related factors such as income and health status (OECD, 2011a; Blanchflower and Oswald, 2008; Deaton, 2010). However, among lower- and middle-income countries, there is evidence to suggest that life satisfaction decreases with age, and age-related decreases appear to be particularly marked in transition countries in Eastern Europe and the former Soviet Union (Deaton, 2010). Gender also has a small but significant impact on reported subjective well-being in several countries, as detailed in the preceding section. Finally, there may be systematic differences in patterns of responses associated with different cultures, discussed in the section that follows.

When making group comparisons in subjective well-being, it is therefore important that the gender, age and cultural composition of the groups in question are taken into account. This is also important when examining changes in national subjective well-being over time, which could be linked to demographic shifts among the population. Without taking these factors into account, spurious interpretations of the data are possible. For example, whilst it is evident that the onset of widowhood is associated with a very substantial decrease in life evaluations at the individual level (Clark and Oswald, 2002), the probability of widowhood increases with old age, and in affluent countries life evaluations also tend to increase after the age of 55. This means that in cross-sectional national data, the life evaluations of widowed individuals may not be as low as one might expect, due to the interaction between widowhood risk and old age. In a recent small-scale study in New Zealand (UMR, 2012), for example, widows appeared to report *higher* happiness than married respondents, a pattern that is almost certainly due to widows being, on average, older than married, divorced or single respondents.

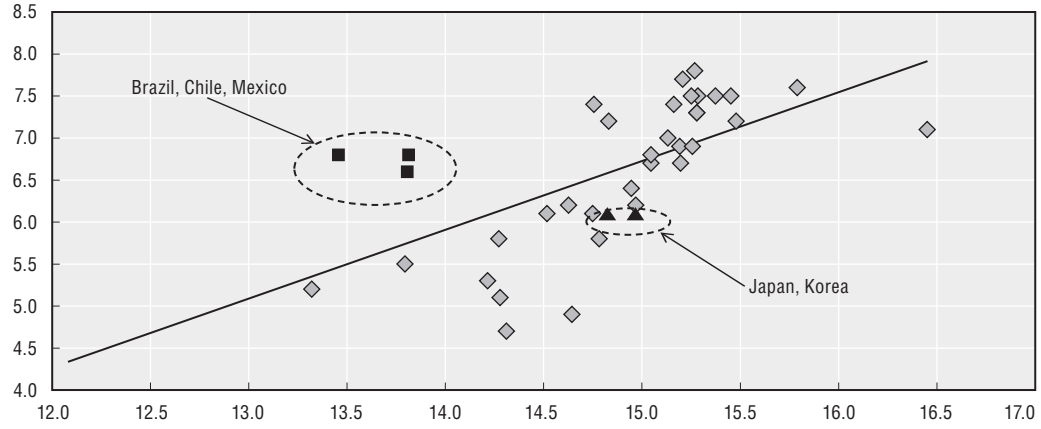
What is the role of culture in international comparisons, and can data be "corrected" for "cultural bias"?

The fact that the global distribution of both life satisfaction and affect is wide and varied suggests that differences in country-level life circumstances are likely to produce differences in country-level subjective well-being, and this has been confirmed empirically in a wide variety of studies based on large international datasets (e.g. Helliwell and Barrington-Leigh, 2010; Helliwell et al., 2010; Helliwell, Layard and Sachs, 2012; Deaton, 2010; Boarini et al., 2012; Fleche, Smith and Sorsa, 2011). For example, in the Cantril Ladder life evaluations data set examined in the *World Happiness Report* (Helliwell et al., 2012), the countries with the top four rankings are reported to have average incomes 40 times higher than those countries with the bottom four rankings.

On the other hand, countries with relatively similar levels of economic development can sometimes report quite different mean levels of subjective well-being.[20] Inglehart, Foa, Peterson and Welzel (2008) illustrate how international measures of subjective well-being data can diverge from the pattern that might be predicted based solely on their level of economic development (see also Figure 4.7). This indicates that Latin American countries

Figure 4.7. **Subjective well-being (SWB) and per capita gross domestic product (GDP)**

Mean life satisfaction[1] versus log of GDP per capita – OECD and selected countries, 2010



1. Mean Cantril Ladder score.

Source: OECD (2011), *How's Life?*, OECD Publishing.

in particular tend to report higher levels of subjective well-being than might be expected based only on their GDP per capita, whilst ex-Communist countries appear to report lower subjective well-being than one might expect.

While income may be an important determinant of country differences in subjective well-being, a number of other non-economic factors are also important, many of which relate to measurable differences in life circumstances (such as health and social context, see Helliwell et al., 2012 and the discussion below). As with interpersonal comparisons of subjective well-being, there may also be some differences *between countries* in terms of the frames of reference used by individuals to report on their own well-being, as well as differences in how questions are understood and interpreted, and how response formats are used. Chapter 2 described the evidence around hypothesised cultural response styles and the methodological steps that can be taken to reduce the risk of differences in how scales are understood by respondents, and Chapter 3 covered the issue of scale translation in more detail. The purpose of the present discussion is therefore to focus on the interpretation of observed international differences in average levels of subjective well-being, possible sources of those differences, and whether data can and should be "corrected" for linguistic or cultural bias *after* it has been collected.

Cultural impacts versus cultural "bias". Before attributing differences in average subjective well-being between countries at similar levels of economic development to "cultural bias", it is important to remember that these differences may have many sources. A helpful distinction can be made between *cultural impact*, which refers to valid sources of variance between cultures, and *cultural bias*, which refers to inter-cultural differences that result from measurement artefacts[21] (Van de Vijver and Poortinga, 1997).

If we assume that an identical measurement approach has been adopted across all countries, and therefore that observed differences cannot be attributed to the methodological differences described in Chapter 2, differences in average levels of subjective well-being between countries may have at least four different sources:

- *Life circumstances*

  In addition to income and other economic variables, there may be other important differences between countries in terms of social context and other life circumstances. As noted above, levels of economic development are just one group of potential drivers of subjective well-being, but a very wide variety of others exist, often playing a more substantial role than income (e.g. health, social relationships, unemployment rates, freedom of choice and control). These drivers can include valid country differences in subjective well-being connected to levels of democracy, tolerance of outgroups, strength of religiosity, or trust in others (Inglehart et al., 2008; Bjørnskov, 2010), and perceived freedom, corruption and the quality of social relationships (Helliwell, 2008; Helliwell et al., 2010; Helliwell et al., 2012). The socio-demographic structure of countries may also contribute to mean differences observed between countries. Because of the very wide range of factors that impact on average levels of subjective well-being, comparing countries on the basis of income alone is insufficient.

- *Differences in how people feel about their life circumstances*

  There may be differences between countries in how people *feel* about their current life circumstances. Many factors may potentially influence how life circumstances are appraised, including an individual's reference group (i.e. frame-of-reference effects, discussed just above), past life experiences, the past or present political and economic situation, the policy environment and the country's religious, cultural and historical roots. These differences may contribute to *appraisal styles* that influence the connection between objective life circumstances and subjective feelings – for example, the degree of optimism or pessimism individuals feel about the future. Rather than representing cultural "bias", these should arguably be regarded as valid sources of difference between countries – because they influence the level of subjective well-being actually experienced by individuals, even if this does not mirror exactly the measures of their objective life circumstances.[22]

- *Language differences that influence scale use*

  Systematic differences between countries may also arise as a result of imperfect translatability of subjective well-being constructs. For example, Veenhoven (2008) has shown differences between the (0-10) numerical ratings that respondents assigned to English and Dutch translations of verbal response categories (*very happy, quite happy, not very happy* and *not at all happy*). In this instance, linguistic differences would produce biases in how people respond to a verbally-labelled scale that bear no relation to how individuals actually feel about their lives – and thus it would be desirable to remove this bias from the data.

- *Cultural response styles or biases*

  There may be country-specific differences in how individuals *report* their feelings, regardless of their actual experiences. For example, a "modesty" or moderate-responding bias might have a downward influence on self-reports, without having a negative impact on private feelings of subjective well-being. Similarly, tendencies towards "extreme responding" (i.e. using scale end-points) or more socially desirable responding could imply

differences in modes of cultural *expression*, rather than substantive differences in the subjective well-being actually experienced. These effects could be described as group differences in scale use, or cultural response styles. If differences in scale use do not represent differences in how people feel about their lives, they can be regarded as a source of bias that it would be desirable to remove.

It is important to distinguish between these four potential sources of country differences, because they have different implications for the *validity* of between-country comparisons, and for the actions one might take to address country-specific differences in subjective well-being. In the case of unmeasured life circumstances, there is a country-specific effect that may or may not be related to culture. In the second case, differences between countries can reflect cultural impact – i.e. differences in how respondents *genuinely feel*, and which would add to the predictive validity of the overall subjective well-being measure (e.g. in its association with future behaviour or health states). One would not necessarily want to *correct* subjective well-being scores for either the first or the second of these country-specific differences. Linguistic differences or cultural response styles, on the other hand, can be expected to add *bias* to the data, reducing its overall validity and predictive ability. In these instances, it would be desirable to find a way to either minimise the problem at source through survey design (Chapter 2) and translation (Chapter 3), or to adjust the data *ex post* to remove the bias and enhance the overall usefulness of the measures.

Methods for examining and "correcting" cultural bias

Counterfactuals. The "counterfactual" approach attempts to isolate the problem of unmeasured life circumstances by examining the extent to which country differences can be explained by a variety of objective outcomes. It involves using information about objective life circumstances to adjust subjective reports so that only the variance in subjective reports that can be explained by objective life circumstances is retained. For example, Jürges (2007) used detailed information about a wide range of objective medical complaints to adjust self-reported health data from 10 European countries so that only variance that could be explained with reference to the objective health indicators was included.

Two fundamental assumptions in the counterfactual approach are that: a) all relevant objective variables have been included in analyses; and that b) there is no valid variation among respondents in how objective states are perceived and experienced. Using such a procedure to "correct" for country-specific effects in practice eliminates *any* country differences in the relationship between objective outcomes and subjective experiences. In the work of Jürges (2007), this means that any country-specific differences in, for example, the care received by patients, or the support received from friends and family, are eliminated from the adjusted data. Unfortunately, these country-specific differences are precisely the kinds of differences that are likely to be the focus of interest of governments and public service professionals. Thus, in removing the country-specific influences that might "bias" self-reported data, the study design also potentially removes any *substantive* differences between countries in how health conditions affect perceptions of health. However, such an approach may have value in helping to understand the nature and composition of differences between countries (Fleche, Smith and Sorsa, 2011).

Fixed effects models. One could argue that if counterfactuals are to be constructed to better understand subjective well-being differences between countries, they will need to take a wide variety of life circumstances into account – including social context variables that often rely on self-report measures. The difficulty is, however, that cultural differences in scale use are likely to affect many self-report measures, not just subjective well-being. Thus adding self-reported measures into a counterfactual model could serve to mask cultural response styles, rather than teasing them out. An alternative approach involves examining country and regional fixed effects in the data.

Helliwell, Barrington-Leigh, Harris and Huang (2010) reported that among a global sample of between 50 000 and 140 000 respondents in 125 countries, a regression equation including demographics, income, being unable to afford food, having friends to count on, perceived freedom, perceived corruption, charitable donations of money and time, helping strangers, and religion explained between 30 and 45% of the individual-level variance in life evaluations. Adding dummy variables for regional fixed effects added between 0.3 to 1.1% to the variance explained, with a significant and positive coefficient for the South and Central America region (indicating more positive evaluations of life than might be predicted from other variables in the model).[23] Adding individual dummies for every country surveyed added between 2.5 and 4% to the overall variance explained. Importantly, however, neither set of dummies substantially reduced the coefficients for the other predictors in the model – indicating that the strong relationships between social indicators and subjective well-being was not due to country- or region-specific fixed effects. Helliwell et al. (2010) also report that when separate regression models are created for each of the different regions, and each of the different countries, the coefficients are markedly similar to those obtained in the single global model.

The work of Helliwell et al. (2010) suggests that even if some variation in average scores between countries may be due to unexplained factors, there do not appear to be big country or regional differences in the *structure* of the relationship between subjective well-being and some of its key known determinants. Although they did not explicitly set out to measure cultural bias, their results also suggest that when a much larger predictor set is examined, relatively little of the unexplained variance in life evaluations is due to region- and country-specific fixed effects. This does not *preclude* the presence of cultural biases in the data, but if these biases operate at the regional- or country-level, they appear to explain only a small amount of variation in individual-level responding, above and beyond life circumstances.

Vignettes. The "vignette" approach (e.g. King et al., 2004) attempts to measure the different ways in which individuals and/or cultures may understand, interpret, benchmark or respond to the same survey question, issues that are collectively known as "differential item functioning" (DIF). DIF can result from either scale translation problems or cultural response styles.

Vignettes are short descriptions of hypothetical scenarios that respondents are asked to rate, using the same scale format used to obtain self-reports. The vignette method works on the assumption of "vignette equivalence", i.e. that, because respondents are each evaluating *the same* vignette, they should in principle assign identical ratings to that vignette. Thus, any differences between individuals (or groups of individuals) in the ratings assigned to vignettes are attributed to differential item functioning or response styles[24] (often interpreted as "cultural bias" in cross-country studies).

The vignette approach has been used in recent cross-country studies to identify cultural effects in subjective data (Angelini et al., 2011; Kapteyn, Smith and van Soest, 2009; Kristensen and Johansson, 2008 – Box 4.4). For example, Angelini et al. (2011) utilised data from respondents in ten European countries (*N* = 5 606), who were asked to provide life satisfaction ratings for both themselves and for two fictional elderly characters described in two separate vignettes, detailing the characters' age, family and social relations, income and health circumstances. In Angelini et al.'s study, the (unadjusted) self-assessments of life satisfaction showed Danes to be the most satisfied with their lives and Italians the least satisfied. However, the vignette method indicated differences in the scale thresholds used by Danish and Italian respondents to define the response categories (on a 5-point verbally-labelled scale, ranging from "very satisfied" to "very dissatisfied"). In simulations estimating life satisfaction for other countries using Danish scale thresholds, more than 95% of respondents in all countries would rate themselves as satisfied or very satisfied with their own life. In simulations using Italian thresholds, this reduces to between 80 and 95% in most cases, but around 60% for Poland and the Czech Republic.

Box 4.4. **Use of vignettes to investigate job satisfaction – Kristensen and Johansson, 2008**

Kristensen and Johansson (2008) examined subjective assessments of job satisfaction across seven EU countries (*N* = 5 988), using 19 different sets of five vignettes. Respondents evaluated both their own jobs, and fictional jobs described in vignettes, on a 0-10 job satisfaction measure.

As can be seen from Table 4.4, the relative country rankings in terms of job satisfaction shift: a) when objective job characteristics are controlled; and b) when satisfaction rankings are adjusted on the basis of vignette responses.

Table 4.4. **Differences in country rankings of job satisfaction, 2008**

Rank	"Positivity" of vignette scores[1]	Average unadjusted self-report job satisfaction	a) Job satisfaction rankings, controlling for objective job characteristics	b) Vignette-adjusted job satisfaction ranking[2]
1	Finland	Denmark (7.5)	Finland	Netherlands
2	Spain	Finland (7.4)	Denmark	Greece
3	Greece	Netherlands (7.3)	Greece	Denmark
4	Netherlands	Greece (6.9)	Netherlands	Finland
5	United Kingdom	France (6.6)	Spain	France
6	Denmark	Spain (6.5)	France	United Kingdom
7	France	United Kingdom (6.4)	United Kingdom	Spain

1. Calculated from the information provided by Kristensen and Johansson in Table 3, p. 104.
2. Also with objective job characteristics controlled.

Kapteyn, Smith and van Soest (2009) used vignette ratings as a means of examining response scale differences between Dutch and US respondents, and found that Dutch life evaluations become more spread out when re-scaled using US threshold parameters – such that there are increases in the percentage of respondents who are very satisfied (from 21.5% to 26.7%) or very dissatisfied (from 1.5% to 3.1%). However, this re-scaling has little impact on the total scale mean: if one re-calculates mean scores by assigning 1-5 scale

values (from very dissatisfied to very satisfied respectively) the original Dutch ratings average 4.06, and when re-scaled according to US thresholds the mean average is 4.07. From either perspective, then, differential item functioning appears to produce only small scale shifts among this sample.

Although vignette-based studies have revealed some differences between countries in the subjective well-being ratings assigned to vignettes, the *source* of these country differences requires further investigation. The vignette method assumes that any country-specific differences in the scores assigned to vignettes must be brought about by response styles, or cultural bias. Another likely possibility is that this method picks up problems in scale translation between languages, particularly where short and verbally-labelled response categories (typical of vignette studies) are used. However, it is also possible that differences between vignette ratings reflect meaningful differences between countries (such as differences in the work, social and policy environments) that have real consequences for how individuals perceive and respond to the narrow set of life circumstances described in vignettes.[25] For example, estimates of the likely subjective well-being impact of unemployment in a vignette may be affected by country-specific differences in the macroeconomic and jobs climate, the social safety nets available, social norms around work (Stutzer and Lalive, 2004), and total levels of unemployment (Clark, 2003).

Vignettes also require that individuals are able to accurately forecast how they would feel in different circumstances and that people respond to vignettes in the same way that they do to actual questions about their own subjective well-being. Both of these assumptions can be questioned. It is important, therefore, to *empirically demonstrate* that the vignette-adjusted ratings are more accurate reflections of what respondents are subjectively feeling, or more accurate predictors of future behaviour, than are unadjusted mean scores.

Migrant data. Senik (2011) proposes a technique in which the effects of culture can be investigated by comparing the subjective well-being of native and migrant respondents within a country. When a variety of background and economic variables are controlled for, if migrants systematically differ from natives in their subjective well-being assessments (despite exposure to similar within-country conditions), one might infer cultural causes for these differences – although it will be important to eliminate a variety of other ways in which migrants and natives differ in their experiences. It is also possible to compare the subjective well-being of individuals living in their native country with that of their fellow countrymen who have emigrated overseas (e.g. French people living in France, versus French people living in other countries).

In examining the "French unhappiness puzzle", Senik found that across a sample of 13 European countries in total, living in France reduced average happiness by 0.23 points (on a 0-10 scale) and reduced the probability of reporting 7 or more by 19% relative to the rest of the sample. By contrast, people living in Denmark were 50% more likely to score 7 or more. Furthermore, although the general trend across the sample was to find that natives are happier than immigrants, among the French sample this pattern was reversed. The actual reported level of happiness among French natives was 7.22, but when simulated using regression parameters obtained from immigrants living in France, this was predicted to rise to 7.36. Conversely, the average reported happiness of immigrants (7.25 in the original self-reports) was predicted to decrease to 7.15 when simulated using regression

parameters based on responses from French natives. Finally, when estimated using the parameters obtained for natives among the other 12 European countries, the happiness of French natives was predicted to rise to 7.54.

This method thus implies something unique to French natives that has a tendency to depress happiness reports by between one-eighth and one-quarter of a scale point, on a 0-10 scale, relative to immigrants from a wide range of countries, and natives among other European countries.[26] What is much less clear from this work is the precise source of this effect. Senik is clear that international differences identified through her methods should not be interpreted as "meaningless anchoring biases and measurement errors, but as identity and cultural traits" (p. 7). This approach thus views culture and mental attitudes not as something that should be statistically controlled in cross-country comparisons, but rather as genuine cultural impacts which can point to possibilities for policy interventions, particularly in school and childhood experiences, that can support the development of more positive subjective well-being overall. Senik also notes that the observed "French unhappiness" is mirrored by low levels of trust in the market and in other people.

Comparison of life evaluations and affect balance. Other insights into response biases can potentially be obtained through comparing life evaluations and affect balance measures. Recent experiences of affect are thought to be less susceptible to retrospective recall biases than life evaluations in particular (Kahneman and Riis, 2005; Diener and Tov, 2012; Oishi, 2002; Tsai, Knutson and Fung, 2006; Blanchflower, 2009). In terms of response styles or biases that have been linked with culture, such as more extreme and more moderate responding, it is also possible that *affect balance* measures, by subtracting mean negative affect from mean positive affect, reduce the impact of such biases in the final data set.[27] This requires further examination.

Krueger et al. (2009) reported that, based on results from studies with US and French samples, the French (on average) report spending more time in a more positive mood, and spend more of their time in activities that are rated as more enjoyable. This contrasts with responses to life satisfaction questions, which typically find US samples reporting higher life satisfaction than the French, and suggests affect data can add something new to the overall well-being picture.

In an analysis of subjective well-being among 40 OECD and emerging countries, there are some marked differences between affect balance and life satisfaction measures in terms of countries' relative rankings (OECD, 2011a). For example, Japan falls below OECD-average levels on life satisfaction, but Japan's relative position on affect balance almost reverses, such that it is ranked third-highest overall. All Asian countries considered in this data set ranked higher using affect balance relative to life satisfaction, and this was particularly striking for China, Indonesia and Japan, who move from near the bottom of the rankings to near the top. In contrast, Israel, Italy, Finland, Switzerland and Canada ranked significantly lower on affect balance relative to life satisfaction.

However, the fact that the two measures perform differently doesn't in any way show that one is less "biased" than the other – and there are several alternative explanations for observed differences in the patterns among affect and life evaluations. For example, it is known that these measures are affected in different ways by objective life circumstances, such as income, which has a stronger impact on life satisfaction than on affect balance (Boarini et al., 2012;

Kahneman and Deaton, 2010). As a result, affect data could not be used to "correct" life evaluations, but comparisons of the two measures are nonetheless interesting in the context of other investigations of cultural impact – including vignettes and migrant data.

Conclusions: What is the role of culture in international comparisons, and can data be "corrected" for "cultural bias"?. The gap between standard economic variables and subjective well-being arguably reflects where using subjective well-being can add value to existing measures of progress. However, this gap contains noise as well as signal. Separating the signal from the noise is a particularly vexed issue when it comes to comparisons of average subjective well-being levels between countries.

While counterfactuals, vignettes, migrant data and comparisons between life evaluations and affect offer interesting insights into the impact of country (and, by extension, culture[28]) on subjective well-being, none of these approaches has yet been able to convincingly distinguish between substantive cultural *impacts* and cultural *bias*. The relatively small number of countries sampled in the existing research also makes it difficult to extrapolate more widely on the basis of this work – meaning that there is little that can be said even about the expected *magnitude* of cultural effects, particularly at a global level. For example, although the "cultural differences" between US and Dutch respondents (Kapteyn et al., 2009) or French and other European respondents (Senik, 2011) appeared to be reasonably small, more disparate cultural groups, such as Latin American and former Soviet countries, may reveal larger differences. The findings of Helliwell et al. (2010, cited earlier) suggest that when a wide range of predictors are taken into account in a global sample, both region- and country-specific fixed effects (which are likely to reflect both unmeasured variables *and* sources of systematic biases if these vary between the regions and countries identified) explain a relatively small percentage of the overall variance in subjective well-being (between 0.3 and 4%). Access to further high-quality data on subjective well-being from large and nationally-representative samples will help to shed light on the issue of *what proportion* of average-level differences between countries can be attributed to cultural biases in determining whether the benefits of data adjustments outweigh the costs of lost information.

A further practical limitation in using vignettes and migrant data to "correct" country averages of subjective well-being data is that the impact of culture in the data cannot be quantified in simple absolute terms – rather, it is defined relative to other countries in the sample. This provides a further challenge if the goal is to adjust national-level data to provide "culture-free" estimates, and implies that only a large and representative global sample could really be used as a basis for such adjustments.

Given the current state of the evidence available, these guidelines do not recommend using the methods described in this section for "correcting" mean average country-level subjective well-being data for cultural influences. It is not yet clear, for example, whether adjusting subjective well-being according to these methods actually *adds* to the validity of the data or to its usefulness in terms of predicting future behaviour and other well-being outcomes. Correcting data for *all* country-specific influences on how objective circumstances are perceived would risk removing the influence that all unmeasured country differences (including the influence of a country's policy environment, social safety nets, and a wide range of valid cultural differences) have on how subjective assessments and feelings are formed.

Thus, until further research and analyses become available, the risk of cultural bias is best managed through adopting survey design and translation principles (described in Chapters 2 and 3 of these guidelines) that seek to minimise differences in scale interpretation and use among respondents. Inclusion of covariates that could help to explain cultural *impacts* may also be valuable in international comparisons. Finally, supplementing life evaluation data with affect balance measures may provide a more rounded picture of country differences overall.

2. Better understanding the drivers of subjective well-being

Introduction

The present section is concerned with analyses examining the drivers of subjective well-being. If identifying vulnerable groups and international benchmarking are core elements of monitoring well-being, better understanding the drivers of subjective well-being can help to *explain* some of the differences observed over time or between groups – both within and among nations. This analysis might then suggest areas where policy interventions and individual life choices might raise levels of subjective well-being overall.

This section is organised in three parts. It begins with an overview of what better understanding the drivers of well-being means in practice, and the types of drivers typically examined in such analyses – including other high-level well-being outcomes, life events and more specific policy interventions. The use of subjective well-being data to inform the appraisal, design and evaluation of different policy options, as well as to examine policy trade-offs, is also described.

Key methods involved in the analysis of subjective well-being drivers are then covered in the second part of this section. This includes discussion of data requirements and the types of survey design that facilitate causal interpretations, as well as brief consideration of the types of statistical analysis involved in these investigations. Finally, the third part of this section discusses the challenges associated with interpreting analyses of the drivers of subjective well-being. The fundamental questions addressed are *what size of impact can be expected?* and *how can the impacts of different drivers be compared?* Key issues considered include interpretation of regression coefficients, the generalisability of results, the risk of error in the measurement of both drivers and outcomes, and the time frames under investigation.

What does "better understanding the drivers" mean, and why does it matter?

Understanding the drivers of subjective well-being means identifying variables that appear to have causal relationships with subjective well-being and examining some of the mechanisms through which drivers take their effects. Drivers of subjective well-being can include high-level well-being outcomes, such as income and health conditions, as well as specific life events and circumstances such as unemployment or the onset of disability, or specific patterns of behaviours and time use, such as commuting, watching TV, or interacting with friends and family.

Governments and researchers may be interested in the drivers of subjective well-being for a number of reasons, which are described in more detail in the sections that follow. Organisations and individuals may also have an interest in both the life circumstances and the daily events that influence subjective well-being in order to help inform decision-making and increase the well-being of workers and their families (Box 4.5).

Box 4.5. **Wider public interest in the drivers of subjective well-being**

Drivers of subjective well-being that could interest the general public include:

- Other high-level well-being outcomes (such as income, social connections and health) and the trade-offs that may exist between them.
- The impact of certain life events, and the factors associated with positive adaptation to life events over time.
- How time use plays a role in both short-term mood states and longer-term well-being.

For example, Layard (2005) discusses how geographic labour mobility might bring positive economic benefits, but could potentially lead to an overall decrease in well-being, including subjective well-being, through losses in both work and social connections and the weakening of local community ties. The trade-off between economic benefits and the "hidden costs of mobility" (Dolan and White, 2007) can be explored through examining subjective well-being data, which can also be used to illuminate the factors associated with successful adaptation to relocation. This information may be useful for both the individuals making those trade-offs as well as organisations seeking to support the well-being of staff that have been relocated.

Data obtained through a combination of time-use and survey methods may also prove interesting for individual decision-making. Loewenstein and Ubel (2008) view informing the public about the likely consequences of particular actions as the main way in which affect data should be used. Kahneman and Riis (2005) meanwhile suggest that paying more attention to the allocation of time is one of the more practical ways to improve experienced well-being.

Several studies have provided insights into both subjective well-being gained from individual activities and the net impact that time allocation has on national well-being. For example, Kahneman et al. (2004) conducted an investigation of affect among nearly 1 000 working women in Texas using the Day Reconstruction Method. Among this sample, the three work-related activities (the morning commute, time spent at work, and the evening commute) were associated with the lowest average levels of positive affect balance, whilst intimate relations, socialising after work and eating dinner were associated with the highest average levels of positive affect balance. These authors propose calculating *national accounts of well-being*, based on the proportion of time individuals report being engaged in different activities, and the net affective experience reported during each of those activities (e.g. Krueger et al., 2009).

Individual activities that have been investigated in detail for both their short- and long-term influence on subjective well-being include TV watching, Internet use and commuting. Frey, Benesch and Stutzer (2007) for example reported that people watch more TV than they consider optimal for themselves, and that heavy TV viewers – particularly those with significant opportunity costs of time – report lower life satisfaction. Gross, Juvonen and Gable (2002) examined Internet use among adolescents and found that, whilst the overall duration of time spent online was not associated with evaluative subjective well-being or daily affect, the emotional closeness of instant message communication partners was associated with daily social anxiety and loneliness in school. Finally, Stutzer and Frey (2008) found that people with longer commuting times reported systematically lower satisfaction with life overall – consistent with the finding from Kahneman et al. (2004, above) that commuting is associated with low levels of positive affect balance.

Because of the various challenges associated with interpreting analyses of drivers, however, it may be misleading to place too much emphasis on comparing the relative effect sizes of the different drivers (see below). In communicating with the wider public, then, results of these types of analyses should not be presented as a *recipe* for subjective well-being, but rather more as a list of ingredients, with a broad indication of their impacts – and allowing flexibility to adapt the recipe according to taste.

Life events and life circumstances as drivers of subjective well-being

Examining the relationship between subjective well-being and other key well-being outcomes (e.g. income, jobs, health, social relationships, work-life balance, personal security, education, civic engagement, governance, housing, environmental quality) is then the first step for better understanding differences in subjective well-being observed between different groups or over time. As well as enhancing understanding of well-being as an over-arching construct, analysing the drivers of subjective well-being data offers a way to test empirically whether the outcomes currently used to measure and describe societal progress align with the outcomes that determine people's perceptions of their own well-being. These analyses can also assist in identifying potential opportunities for new policy approaches, improving the design of existing policies or highlighting areas where policies or regulations can be withdrawn in order to improve subjective well-being outcomes. In particular, identifying the factors that influence how people react to adversity, such as the onset of disability or unemployment, as well as how successfully they adapt to these events over time, may be particularly relevant to policy-makers.

Although not all of the life circumstances that are important to subjective well-being will be amenable to policy interventions, subjective well-being evidence can have particular implications for government approaches to issues such as: mental health and resilience; employment, training and labour market flexibility; child welfare, and family and community policies; and taxation approaches to products (e.g. addictive substances) and activities known to have an impact on subjective well-being (Layard 2005; 2011). The next step may then be to examine the impacts that particular interventions are likely to have on subjective well-being outcomes, and how consideration of subjective well-being impacts can be used to inform certain policy design features.

Using subjective well-being data to inform the options appraisal, design and evaluation of policies

When making funding allocation decisions, it is important for governments to have information about the efficiency with which resources can be used to achieve policy objectives. Estimating the efficiency of expenditure, often described as *value for money*, delivered by a project, programme or policy intervention involves quantifying the various impacts it might have on outcomes of interest – including economic outcomes (e.g. does the intervention boost jobs or decrease regulatory burdens on business?), social outcomes (e.g. does the intervention improve educational attainment or health outcomes?) and environmental outcomes (e.g. does the intervention contribute to carbon reduction or increase peoples' access to green space in the local area?). These questions are relevant in the process of initially appraising policy options, but may also be asked as part of ongoing refinements to policy design and implementation, as well as when examining the potential impacts of *stopping* a particular policy intervention or regulation.[29]

Options appraisal takes place before a policy is implemented, whereas policy evaluation involves the assessment of policy during its implementation – and might include both specifically commissioned research, as well as less formal evaluations based on existing evidence. Formal programme evaluations may involve an experimental or quasi-experimental design for investigations and include measures both before and after a policy has been introduced, enabling causal inferences to be drawn about the impact of

the policy. Policy design considerations are relevant both before and during policy implementation, as well as in interpreting the evaluation of policy impacts and what can be done to enhance them.

Alongside some of the typical economic, social and environmental outcomes, subjective well-being data can provide policy-makers with an additional perspective on the potential impact of a policy. As noted earlier, subjective well-being data may offer unique insights into the effects of a given action, taking into account a variety of objective well-being outcomes and how they combine to produce an overall *perception* of well-being. It should be noted, however, that because subjective well-being has so many drivers, the impact of any one policy, particularly one that affects only a small number of people, may prove difficult to detect. This has implications for the sample sizes required, and the study design adopted in, for example, formal policy evaluations – issues that will be discussed in the section that follows.

In assessing the likely impacts of a policy intervention on subjective well-being, analysts are likely to draw on prior literature, including academic sources, and international examples. More comparable data will enhance the quality of these sources of information and provide better baseline information about the levels of subjective well-being to expect among different population sub-groups. This baseline data can provide essential information about the "do nothing" policy option – i.e. what to expect in the absence of intervention.

Diener and Tov (2012) list a wide variety of policy considerations where it may be valuable to consult subjective well-being data. These include issues such as deciding how to support day care for elderly Alzheimer patients, examining the moods and emotions of caregivers when the patient is in day care or at home and the life satisfaction of caregivers when respite care is provided; or examining the well-being benefits of parks and recreation, testing whether parks are more crucial to well-being in areas where dwellings have no outdoor space and whether life satisfaction is higher in cities with plentiful parks than in cities where parks are rare.

In terms of applied examples, Gruber and Mallainathan (2002 have used subjective well-being data to examine optimal cigarette taxation across a range of areas in the United States and Canada. Boarini et al. (2012) show how data from OECD member countries can be used to explore the impact of health co-payments and unemployment replacement rates on national levels of subjective well-being, as well as well-being among certain population subgroups, such as people working versus those outside the labour market. Using a quasi-experimental design, Dolan and Metcalfe (2008) examined the subjective well-being impact of an urban regeneration project in Wales.

Using subjective well-being data to inform policy trade-offs

A key part of governing involves making decisions not only between different policy options but also between different policy objectives. When confronting policy *trade-offs*, one of the perennial challenges lies in comparing the relative value of different economic, social or environmental policy outcomes. This is like comparing apples with oranges: is an increase in educational attainment any "better" or "more necessary" than an increase in health outcomes? Governments may use various methods to assist in making these decisions, such as international benchmarking and national targets, but policy trade-offs, and particularly those involving different departments with different objectives, remain thorny issues to resolve.

Although it is difficult to imagine ever solving this problem through evidence alone, subjective well-being data offer one way of looking at the *societal preferences* (Loewenstein and Ubel, 2008) for different trade-offs. In the terms of Bjørnskov (2012), governments face a *massive information problem* – i.e. in a large society it becomes impossible to know enough about the detailed preferences and needs of the population to direct policies according to those preferences and needs. By providing information about what is likely to increase the subjective well-being of the population at large, subjective well-being data offers an alternative to listening to the arguments of relatively narrow interest and lobby groups. Subjective well-being can also offer a standard unit of comparison, thereby facilitating more "joined-up government" where departments are better able to consider the spill-overs from their interventions onto a wider range of domains (Dolan and White, 2007).

Practical examples of trade-offs that have been examined in the literature include the trade-off between inflation and unemployment (Di Tella, MacCulloch and Oswald, 2001) and that between income and airport noise (Van Praag and Baarsma, 2005). Fitoussi and Stiglitz (2011) highlight the value in further investigating the well-being impacts of moving towards greater flexibility in the labour market: whilst labour market flexibility is assumed to deliver strong economic benefits, it could also negatively affect two key determinants of well-being, i.e. the quality of jobs and economic security. Although Shackleton (2012) warns about the risks of adopting a "one-size-fits-all" approach to well-being at work, particularly across different countries, having better data with which to examine these trade-offs is surely important.

As noted earlier, attempts to directly compare the different drivers of subjective well-being face a number of challenges, and there will be limits to the extent that these analyses contain *answers* to policy trade-offs *per se*. Some of the interpretive challenges associated with analyses of the drivers of subjective well-being – and particularly the issue of comparing different drivers – are discussed in more detail in the interpretation section that follows.

Methods: How can drivers of subjective well-being be analysed?

Data requirements, survey design principles and causality

Research aimed at better understanding the drivers of subjective well-being requires the inclusion of a wide range of co-variates in the analyses, including a number of standard demographic and control variables (described in Chapter 3), as well as measures of the drivers of interest, and their potential co-variates. Examples of key variables of interest are outlined in Diener, Diener and Diener (1995), Dolan, Peasgood and White (2007), Fleche, Smith and Sorsa (2011), and Boarini et al. (2012). Analyses of drivers require access to micro-level data, and may be undertaken by government analysts, researchers and organisations or institutes with an interest in informing government policy, academic enquiry and public discourse, as well as in organisational well-being (including business approaches to employee well-being).

Ideally, research into the drivers of subjective well-being should utilise data that enable some inferences about the causality of relationships between variables. Although the gold standard for determining causality involves experimental manipulations of hypothesised drivers under controlled conditions, this is close to impossible for most of the policy-relevant determinants of life evaluations and eudaimonia in particular. The model scenario for determining causality in real-life settings therefore tends to be *randomised controlled trials* (RCTs), which involve the random allocation of individuals into groups, each

of which are assessed before and after receiving a different treatment (e.g. one group receives intervention A, one group intervention B, and a third group acts as the control, receiving no intervention). In practice, for many of the potential policy drivers of subjective well-being, RCTs are also rare, particularly in terms of ensuring perfect randomisation and double-blind or single-blind conditions.[30]

Quasi-experimental designs refer to "natural experiments" where a group of respondents exposed to a particular intervention can be matched with and compared to a similar group of respondents not exposed to that intervention. However, investigators tend to have little control over either the level of variation in the determinant of interest, or in the allocation of treatment groups, which is rarely completely random. It may also be difficult to obtain pre-intervention outcome measures – thus inferences may have to be based on measures collected only after the event has occurred. However, quasi-experimental designs do offer some advantages over RCTs, particularly where it would be unethical to randomise treatments, and/or where experimental designs can be challenged in terms of their real-world applicability.[31] Quasi-experimental designs can also enable researchers to draw on much larger and more representative data sets, including national-level data.

Quasi-experimental designs typically require panel data (longitudinal surveys collecting repeated measures of individuals over time) or the collection of pre- and post-data for the populations of interest. These data offer the opportunity to explore whether a change in the level of a given determinant is associated with a subsequent change in subjective well-being over time.[32] While panel data do not enable the researcher to experimentally manipulate the main variables of interest, and panels can suffer from attrition, this approach has the benefit of being able to utilise data sets from large and high-quality samples such as those obtained by national statistical agencies – thus enhancing the representativeness of the sample, and the generalisability of the findings.

Large sample sizes are particularly important for detecting the impact of minor drivers and/or drivers that typically affect only a small proportion of the overall population. In comparison to more experimental methods (such as RCTs), observational data also carry less risk of experimental demand characteristics (e.g. Hawthorne or placebo effects), where a respondent's knowledge that he or she is part of a special treatment group may influence subjective well-being outcomes and/or how they are reported. The same is true of international comparisons, which offer another form of natural experiment, where a particular intervention has been applied in one country but not in another. However, it is very difficult to infer causality from international comparisons of cross-sectional (rather than longitudinal) data, given the variety of uncontrolled differences between countries in terms of both sample characteristics and other variables of interest.

At present, the majority of studies investigating the drivers of subjective well-being tend to rely on cross-sectional data, simply because these are the most widely-available data sets. Strictly speaking, these analyses are concerned with *co-variates* rather than *drivers*, although the term drivers will be used in the sections that follow to denote the underlying intention of the analyses described. Cross-sectional data do not enable causal inferences to be made directly, but can be interpreted alongside evidence about the direction of causality from other sources.

Methods for analysis: Tests of association

The most appropriate method for analysis depends largely on the type of data collected, the method of collection and the nature of the research question. The simplest test for the strength of a relationship between two variables is a bivariate correlation. The Pearson or product-moment coefficient can be calculated when the data are assumed to be normally distributed and the expected relationship between them is linear; Spearman's Rank and other non-parametric tests are available for ordinal data and non-linear relationships. Partial correlation enables examination of the relationship between two variables while removing the effect of one or two other variables. Correlations indicate the possible existence of a predictive relationship between two variables, but they do not imply causation.

For more thorough examination of the drivers of subjective well-being in cross-sectional, international and longitudinal studies, regression analysis is widely adopted. Regression is a correlation-based statistical technique that examines how well a set of explanatory or *independent variables* can predict a given *dependent variable*, i.e. the chosen subjective well-being measure. Regression is particularly suited to complex real-life problems because it allows the impact of several independent variables to be assessed simultaneously in one model, and it can tolerate independent variables being correlated with one another. However, the "best" regression solution (in terms of variance explained per independent variable) is produced when each independent variable is strongly correlated with the outcome variable, but uncorrelated with other variables, whether these other variables are included or excluded from the model. If two correlated independent variables are both included in the same regression model, their relationship with the dependent variable may be obscured (Tabachnick and Fidell, 2001). However, if an independent variable is correlated with some other excluded variable with causal claims, then the included variable will falsely be given credit for explanatory power really due to the excluded variable (a difficulty commonly described as the "omitted variable problem").

Given the ordinal nature of subjective well-being measures, linear regression models (based on ordinary least squares estimates) are theoretically inefficient when compared to methods designed to analyse ordinal outcomes (e.g. Probit). However, Ferrer-i-Carbonell and Frijters (2004) have examined both methods in relation to subjective well-being drivers, and concluded that in practice there are few differences between estimates based on ordinary least squares estimates and Probit methods. Similar results were reported by Frey and Stutzer (2000), and reviewing the literature overall, Diener and Tov (2012) reach a similar conclusion. As the interpretation of ordinary least squares outputs is more straightforward, these are often the results reported. However, it remains advisable to examine both Probit and ordinary least squares approaches in the course of analyses to test for (and report on) any major differences between the results observed.

Where curvilinear relationships are expected, such as in the case of both income and age in predicting subjective well-being, squared values (in the case of age, where the expected relationship is U-shaped) and log values (in the case of income, where the expected relationship is asymptotic) are typically used in regression models.

Other analytical options that may be used to investigate the drivers of subjective well-being include structural equation modelling, also known as causal modelling or analysis, analysis of co-variance structures or path analysis. Like regression, structural equation modelling involves examining a set of relationships between one or more independent variables and a dependent variable (or sometimes several dependent variables); but rather

than using raw measured variables as they are observed, structural equation modelling combines factor analysis[33] with regression – and involves multiple regression analysis of factors. The key advantage of this approach is that it enables complex pathways to be tested simultaneously, and by focusing on relationships among underlying factors (rather than measured variables), estimated relationships are thought of as "free" of measurement error.[34] Detailed discussion of structural equation modelling is beyond the scope of this chapter, but because it also involves association- and regression-based techniques, some of the issues raised below remain relevant.

What constitutes a significant association?

Correlation coefficients (here denoted as *r*) range from -1 to +1, with -1 signifying a perfect negative linear association, and +1 signifying a perfect positive linear association. The square of the coefficient (or *r* square) denotes the per cent of the variation in one variable that is related to the variation in the other. Thus, an *r* of 0.60 (*r* squared = 0.36) means that 36% of the variance in the dependent variable is related to the variance in the independent variable. The statistical significance of a given correlation coefficient indicates the likelihood that the coefficient would be found in a sample by chance when no significant association actually exists between the variables.

In regression-based analyses, the overall model "fit" with the observed data is described in terms of the proportion of variance in the dependent variable that is explained by the variance in the independent variables (the overall *multiple-correlation coefficient*, or R^2 value). Statistical significance is used to indicate whether the overall model provides better-than-chance prediction of the dependent variable.

In order to further understand how each independent variable contributes to the prediction of the dependent variable, one examines the set of *regression coefficients* for the independent variables. The size (and sign) of the coefficient for each independent variable indicates how much the dependent variable is expected to increase (if positive) or decrease (if negative) when the independent variable increases by one unit, while holding all the other independent variables constant.

Interpreting drivers of subjective well-being

Two key questions for interpreting analyses of the drivers of subjective well-being are *what size of impact can be expected?* and *how can the impacts of different drivers be compared?*. The first question needs to be considered with reference to the overall sensitivity of subjective well-being measures, as well as the time frames for the analysis. The second question raises issues about the interpretation of regression coefficients – including problems of correlations among independent variables, the effects of unmeasured or omitted variables, the possibility of shared measurement error (or method variance) between variables and outcomes, the possible presence of reverse or two-way causality between variables and outcomes, and the generalisability of results.

What size of impact can be expected?

There are *a priori* reasons not to expect large movements in subjective well-being data as a result of single drivers. Many of the interpretive issues described in the first section of this chapter (looking at basic patterns of change over time and differences between groups) also apply to analyses involving the drivers of subjective well-being.[35] Factors such as the initial distribution of responses, the proportion of the sample affected, the number of

significant drivers in the model, frame-of-reference effects and adaptation can all limit the impact on subjective well-being that one might expect as a result of any one individual driver, as well as the size of sample required in order to detect that impact.

Senik (2011) notes that the typical model R^2 value of an ordinary least squares estimate of life evaluation varies between 3% and 15%, depending on the control variables and drivers included in the model and the sample size. Drawing on two waves of Gallup World Poll data (2009 and 2010) from all 34 OECD countries and including measures for a large number of key known drivers of subjective well-being, Boarini et al. (2012) obtained R^2 values of 0.35 in the case of life evaluations, and 0.19 in the case of affect balance. Fleche, Smith and Sorsa (2011) report cross-country comparisons of key life satisfaction drivers over two to three waves of data (from the World Values Survey 1994-2008) for 32 different countries and find R^2 squared values ranging from 0.40 in New Zealand to 0.14 in Turkey, with an OECD average of 0.22. In Helliwell, Barrington-Leigh, Harris and Huang's (2010) analysis of data from a global sample of between 50 000 and 140 000 respondents in 125 countries, income plus a range of social and cultural variables explained between 30 and 45% of the individual-level variance in life evaluations.

Given the very large number of potential drivers of subjective well-being, these R^2 values suggest that the proportion of variance explained by any one driver is likely to be small. Furthermore, if the initial amount of variability in a given driver is itself limited (for example, because only a small proportion of the sample is affected by it), then the proportion of variability in the subjective well-being outcome it can explain will also be limited. The key statistic of interest in the analysis of drivers, however, is not the overall model R^2, but the size and significance of the individual regression coefficients associated with each driver – which (in the absence of correlations among independent variables) indicate how much the dependent variable is expected to increase or decrease when the independent variable increases by one unit.

As noted in Section 1, a mean change of 0.3 or 0.5 scale points on a 0-10 life evaluation scale may represent a very sizeable result that one might expect only in response to major life events at the individual level. Between countries, differences may be larger due to the cumulative impact of differences across a wide range of subjective well-being drivers.

Using an appropriate time horizon for analysis is another key consideration when interpreting effect sizes. As with any attempt to evaluate the impact of a driver, it will be important to consider the mechanisms through which that driver is assumed to operate – and how long it may take for these effects to emerge. Due to psychological resilience and adaptation over time, the immediate impact on subjective well-being for some life events and interventions may be greater than the impact several years down the line. As noted previously, the variables that influence the rate and extent of adaptation to life events over time may be of key interest to policy-makers, and thus adaptation should be considered a feature rather than a "bug" in subjective well-being data. Nonetheless, the process of psychological adaptation means that close attention to time horizons is warranted, in particular to avoid misinterpreting the effects of exogenous events.

How can the impacts of different drivers be compared?

Particularly in the case of investigating policy trade-offs, or when deciding between two different courses of action, there may be times when it is useful to focus on the *relative size* of different drivers of subjective well-being. Directly *comparing* the regression coefficients

associated with different drivers requires caution, however. Challenges include the need to consider units of measurement as well as the potential for correlation among the drivers, and problems potentially arising due to shared method variance or self-report biases. The generalisability of findings obtained through regression analysis is also of crucial importance for understanding policy implications.

Although the issues associated with comparison of regression coefficients apply to all regression-based analyses, they are particularly relevant here because there are often strong inter-correlations among the drivers of subjective well-being, and because a growing literature suggests some degree of reverse-causality between life evaluations and its drivers in particular. The problems of shared method variance and the possibility of self-report "bias" are also discussed. Finally, a key consideration for policy evaluations in particular are the time horizons over which well-being drivers might be expected to take effect.

Regression coefficients and correlations among independent variables. As noted earlier, when there are no interrelationships between independent variables, the size of the regression coefficient gives an indication of how a one unit change in the independent variable can be expected to influence the dependent variable. However, because the high-level drivers of subjective well-being (such as income, health status and social connections) are likely to be so strongly interrelated, interpretation of their individual contributions must proceed with caution, because there may be *mediation, confounding* and *suppression* effects in the data.

Although conceptually distinct, mediation, confounding and suppression each describe scenarios when relationships between an independent variable and a dependent variable are affected by the presence of a third related independent variable (the *mediator, confound* or *suppressor*). When the third variable is actually measured, these effects can be detected by a substantive change in the regression coefficient for the independent variable when the third variable is included in the model (relative to a model that excludes the third variable). In the case of mediation, the third variable is described as "transmitting" the effect of the independent variable to the dependent variable. In the case of confounding effects, the third variable is described as a "nuisance" variable, producing spurious correlations between the independent variable and the dependent variable.[36] Conversely, when suppression effects are present, relationships between the independent and the dependent variable become *stronger* when the third variable is included in the model. In the event that the third variable remains unmeasured, the coefficient observed for the independent variable can be misleading (see Dolan, Peasgood and White, 2007).

Boarini et al. (2012) also raise the possibility of "over-measurement" of individual drivers – where, if several measures of the same driver are included, correlations among the measures can crowd one another out, such that some relevant variables fail to reach significance in the overall model. This means that a significant driver could be overlooked if there are too many measures of it in the model, because the overall effect will be distributed among too many independent variables.

The fact that the regression coefficient for an independent variable is often dependent on the other variables in the regression equation means that selecting the variables to include in an analysis is a crucially important task. A clear theoretical structure and an understanding of the hypothesised causal pathways must underpin these decisions. While the use of hierarchical (or sequential) regression and structural equation modeling can

provide an analytical strategy for examining causal pathways among variables, techniques of this sort cannot provide a definitive solution with regard to the relative "importance" of interrelated independent variables, in terms of absolute size of impact.

The omitted variable problem. In addition to the mediating, confounding and suppression that can occur as a result of measured variables, causal inferences about relationships between variables can be severely hampered by *unmeasured* or "omitted" variables. Specifically, a significant *statistical* relationship can be observed between two variables not because there is a causal relationship between them, but because both variables are causally related to a third unobserved variable that has been omitted from the analyses.[37] Omitted variables can also suppress causal relationships between observed variables, causing results to fail to reach statistical significance due to unmeasured factors.[38] This is a problem right across econometric analyses (and all tests of association) and is in no way limited to examination of the drivers of subjective well-being.

Among subjective well-being data sets, because so many drivers help to explain final subjective well-being outcomes, several counter-intuitive findings (such as repeated failures to find relationships between income growth and subjective well-being, despite strong cross-sectional relationships between income and well-being) could potentially be explained with reference to variables that have been omitted from analyses (such as changes in *relative* income, or patterns of decline in other important determinants of subjective well-being, such as health, social connections, perceived freedom, corruption, etc.). The effects of relative income, and aspirations about income, have in particular been studied by several authors (for reviews, see Dolan, Peasgood and White, 2007, and Clark, Frijters and Shields, 2008) and reflect the *frame-of-reference effects* discussed in Section 1 of this chapter.

Another set of omitted variables often discussed in relation to subjective well-being involve personality- and temperament-based measures. Individual fixed effects do appear to account for a sizeable proportion of the variance in subjective well-being measures (see Diener, Inglehart and Tay, 2012), and some of these fixed effects may in turn reflect dispositional tendencies. For example, Lucas and Donnellan (2011) reported that 34-38% of variance in life satisfaction was due to stable trait-like differences – although this study did not include measures of the objective life circumstances that might impact on stable trait-like components. The issue of whether these stable differences reflect a true causal impact of personality and temperament on experienced subjective well-being, or simply a response style that affects self-reported measures (including subjective well-being, but also health and exposure to stress) is discussed in relation to shared method variance, below.

Self-report measures and shared method variance. A final factor to consider in the interpretation of subjective well-being drivers is shared method variance,[39] also known as common method variance, which can inflate the estimated impact of self-reported drivers relative to those measured through other means (such as objective observations). In particular, due to a combination of social desirability biases, response sets, differences in scale interpretation or use, and similarities between the questions themselves, one might expect that subjective well-being and other self-report measures such as self-rated health, self-reported mental health, self-reported social connections, and/or personality and dispositional variables might have correlated errors. Indeed, response formats to such questions are often very similar (e.g. 0-10 labelled scales). Furthermore, several items on

current measures of eudaimonia and affect bear a strong resemblance to some of the questions used to measure personality and mental health. Concepts such as self-efficacy, often included in constructions of eudaimonia, are also often considered to be an aspect of personality or dispositional tendency.

When comparing the effects of different drivers of subjective well-being, particularly in cross-sectional data, it is therefore important to consider *how* each of those determinants was measured. The possibility of shared method variance has led some authors to suggest that dispositional measures such as personality or negative affectivity[40] should be included as control variables in analyses of self-report data, and particularly when analyses are cross-sectional (Brief et al., 1988; Burke, Brief and George, 1993; McCrae, 1990; Schaubroeck, Ganster and Fox, 1992), in order to remove any bias associated with subjective self-report processes more generally. The risk in doing so, however, is that this 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. Personality, and negative affectivity in particular, may also play a substantive role in the overall development of subjective well-being (Moyle, 1995; Bolger and Zuckerman, 1995; Spector et al., 2000).

An alternative approach to investigating the issue of shared method variance and self-report bias is to use longitudinal panel data, in which individual fixed effects can be controlled. In such models, the ability of self-reported drivers to predict changes in subjective well-being over time can be investigated, and this is a much stronger test of causality. These types of analyses enable the effects of more objective indicators to rise to the forefront, whilst problems associated with shared method variance recede into the background.

Reverse and two-way causality. Understanding the direction of causality when examining drivers of subjective well-being is crucial to establishing their policy-relevance. As noted in the *data requirements* section above, an analyst's ability to make causal inferences is strongest where experimental data, or data from randomised controlled trials, is available. Quasi-experimental designs and longitudinal panel data can also offer insights into likely causal relationships, because analyses can be restricted to factors that temporally precede changes in subjective well-being over time. In cross-sectional data, the ability to make causal inferences is severely limited – and thus results need to be interpreted alongside evidence about the direction of causality from other sources.

In regression-based analyses, one method for exploring issues of reverse-causality is to include an *instrumental variable*.[41] Instrumental variables are sometimes used when there are problems of endogeneity in regression models – i.e. when the independent variable of interest is correlated with the model error term. Two-way or reverse causality can be a key source of endogeneity, as can omitted variables (described above). Dolan and Metcalfe (2008) and Powdthavee (2010) report using instrumental variables to obtain better estimates of the exogenous effect of income on life evaluations. This typically increases the estimate of the income coefficient (Fujiwara and Campbell, 2011). In practice, however, it is very difficult to identify appropriate instrumental variables for income, as most of the key variables strongly associated with income also tend to be associated with life satisfaction.

Generalisability of results. Analyses of drivers are strongly affected by both the variables included in the model and the population sampled – which in turn both influence the extent to which results can be generalised. The importance of different drivers of subjective well-being may vary systematically according to certain group characteristics, because different groups within and across societies may be characterised by very distinct initial resource endowments. For example, Boarini et al. (2012) examined the determinants of life satisfaction among different population sub-groups (i.e. by gender, age and participation in the labour market) across 34 OECD countries. While the overall pattern of coefficients was quite similar, there were a number of non-trivial differences in the subjective well-being functions[42] observed in the different groups.

This evidence suggests that, for different population sub-groups, the relative impact of the determinants of subjective well-being may differ. Heterogeneity in the relative size and significance of the drivers of subjective well-being has implications for how we might inform the public about the relative importance of the different drivers. Policies aimed at increasing subjective well-being may also need to consider the distribution of well-being resource endowments among different population sub-groups. Regression analyses generate results for the *average* individual – and in practice, there may be wide individual differences in the specifics of the well-being function. Different people may find happiness in different ways.

Although several studies have highlighted strong consistencies among affluent countries in terms of the direction and significance of some of the high-level determinants of subjective well-being (Helliwell and Barrington-Leigh, 2010; Fleche, Smith and Sorsa, 2011), one might also expect to see some differences in subjective well-being functions between countries, because countries also vary in terms of both their initial resource endowments and how those resources are distributed. For example, Inglehart et al. (2008) found that among less economically-developed countries, there were stronger associations between happiness and in-group solidarity, religiosity and national pride, whereas at higher levels of economic security, free choice becomes a more important predictor. Drawing on data from the Gallup World Poll, Bjørnskov (2010) reports that Cantril Ladder life evaluations showed a strong relationship with levels of GDP per capita among countries with lower relative incomes, whereas social trust became a strong and significant determinant only among countries with higher relative incomes. In the same vein, Helliwell and Barrington-Leigh (2010) show that coefficients on a number of social variables are higher in OECD than in non-OECD countries, while the coefficients on log income were identical in the two parts of the global sample.

3. Subjective well-being as an input to cost-benefit analysis

Introduction

The first two sections of this chapter have largely been concerned with analyses in which subjective well-being is the ultimate outcome of interest. But in addition to the intrinsic value of knowing more about subjective well-being, subjective well-being data can play an important role as an input for other analyses – offering insights into human behaviour and decision-making, as well as on how other well-being outcomes develop (Box 4.6). Thanks, in part, to these kinds of insights, subjective well-being data has also been suggested as an alternative means for estimating the monetary value of non-market factors (i.e. goods and services that do not have market prices) for the purposes of cost-benefit analysis.

Box 4.6. **Subjective well-being insights into health, human behaviour and decision-making**

Beyond the intrinsic value of subjective well-being, evidence suggests it is also important to other aspects of human functioning. Fujiwara and Campbell (2011) summarise a broad range of evidence suggesting that individuals with higher levels of subjective well-being are more likely to get married, earn more money and be healthier. Positive affect, negative affect and measures of life evaluations are associated with better long-term health and greater longevity (Danner, Snowdon and Friesen, 2001; Ostir et al., 2001), as well as shorter-term cardiovascular and immune system functioning (Cohen et al., 2003; Kiecolt-Glaser et al., 2002; Steptoe, Wardle and Marmot, 2005) which may mediate longer-term relationships between emotions and health. For example, Pressman and Cohen (2005) report that people with high positive affect have been shown to be less likely to become ill when exposed to a cold virus, and more likely to recover quickly.

In addition to income and health, subjective well-being may have other implications for economic performance and overall well-being. Research has found prospective links between positive emotions and workplace performance ratings and productivity (Diener et al., 2002; Estrada, Isen and Young, 1997; Wright and Staw, 1999). Keyes (2006) also reports evidence that mentally healthy individuals missed fewer days of work, were more productive at work, and had fewer limitations in daily activities. Summarising existing evidence, Clark and Oswald (2002) report that measures of subjective well-being have been shown to predict the likelihood of job quits, absenteeism and non-productive work, as well as the duration of unemployment. Bertrand and Mullainathan (2001) also found that job satisfaction was a strong predictor of the probability of changing jobs in the future.

Finally, subjective well-being data can offer insights into people's (in)ability to estimate the well-being impacts of both market and non-market factors as well as of life events – enabling us to compare estimates of what makes us happy against the level of happiness actually attained as a result. This work suggests that our ability to predict future well-being gains or losses (or our *affective forecasting*) is subject to various biases, such as irrelevant cues* (Sugden, 2005), lack of sensitivity to the size of the good or service valued (Kahneman and Tversky, 2000; Sugden, 2005), and focusing illusions, whereby "nothing that you focus on will make as much difference as you think" (Schkade and Kahneman, 1998; Wilson and Gilbert, 2005). These biases can produce marked discrepancies between the degree to which we think we want something, and the degree to which evidence suggests it will actually make us happy (described by Gilbert and Wilson, 2000, as "miswanting"). These findings have practical consequences for current methods of cost-benefit analysis (see below), as well as more general relevance, providing individuals with better information about the correlates of subjective well-being, so that they can make more informed choices in their own pursuit of happiness.

* An example is the starting point bias, whereby questions that begin: "Would you pay $x for…?" can heavily bias responses towards the valuation of *x* used in the first question (Sugden, 2005).

This section focuses on how subjective well-being data can complement existing approaches to the valuation of non-market factors. It begins with a brief description of cost-benefit analysis and of why it is useful to place a monetary value on non-market factors. It then describes methods currently used for estimating the monetary value of non-market factors, and the ways in which subjective data may be able to complement these methods. Finally, the section briefly discusses some interpretive challenges and the caveats that need to be applied to valuations obtained through the use of subjective well-being data.

What is cost-benefit analysis, and how can subjective well-being data help?

Cost-benefit analysis (CBA) is one of the tools that governments and organisations often use to inform decision-making about complex social choices that include a variety of different well-being outcomes, including economic, social and environmental outcomes. CBA involves quantifying a number of the foreseen costs and benefits associated with a particular project or policy intervention. The information from this analysis can then be used as part of a wider decision-making process, which may involve a variety of other information sources, such as the results of public consultations and/or the net costs and benefits associated with alternate policy or programme options competing for the same funding. The key measurement challenge in CBA concerns finding methods to adequately value all potential costs and benefits, based on a common metric. As used by economists, the common denominator of choice is usually a monetary value, and the costs and benefits investigated tend to focus on those with established market values.

Approaches to valuation

Where costs (and/or benefits) have explicit economic values observable in the marketplace, these values can be used in the estimation process – although as market prices often fail in some ways, adjustments may be necessary. However, for non-market factors, alternative valuation methods are required to estimate monetary value. *Revealed preference* techniques involve calculating shadow prices, inferred from observed behaviour. *Stated preference* techniques on the other hand involve surveying respondents about their "willingness to pay" in order to gain or avoid a certain outcome, and/or their "willingness to accept" compensation to give up a good or put up with something undesirable.

Both preference-based techniques make the assumption that people make choices on the basis of what will maximise their future well-being, and this will be directly revealed by their patterns of expenditure. Despite numerous challenges to these assumptions, preference-based methods have become standard practice for public policy appraisal in the United States and the United Kingdom (Dolan and Metcalfe, 2008). Sugden (2005) also highlights that public policy-making requires *some way* of accounting for the impact of non-market factors on well-being if policies options are to be rationally compared. This is essential whether or not preference-based approaches, or even CBA itself, are deemed to be the correct methods by which to proceed.

The role of subjective well-being data in valuation

All preference-based approaches rely on people's ability to make rational and accurate judgements about how something will make them feel in future. This is also true of market prices. However, evidence from psychology and behavioural economics suggests that people's rationality may be bounded at best and "coherently arbitrary" (Ariely, Loewenstein and Prelec, 2003) at worst. In particular, various biases have been identified that distort estimates of well-being gained from various experiences (Sugden, 2005; Kahneman and Tversky, 2000; Sugden, 2005; Schkade and Kahneman, 1998; Wilson and Gilbert, 2005; see Box 4.6). Fujiwara and Campbell (2011), Sugden (2005) and Frey, Luechinger and Stutzer (2004) review how these biases challenge stated preference-based approaches to the valuation of non-market factors.

An alternative approach to valuation involves using life evaluations (usually life satisfaction) data to directly estimate the impact of a particular outcome on how people feel after the event – thus replacing a hypothetical judgement with *ex post* calculations of impact based on the level of subjective well-being actually achieved. Relative to stated preference approaches, this should remove problems associated with, for example, focusing illusions, because respondents are not prompted to think about the *source* of their well-being. There is also less risk of strategic responding on the part of individuals, i.e. intentionally over- or under-estimating the value of a good due to personal interests in the outcome of a valuation process.

Methods: How can subjective well-being data be used to value non-market factors?

Clark and Oswald (2002) suggest a method by which the life satisfaction gained or lost from experiencing certain life events can be converted into a monetary figure. This is done by estimating the life satisfaction gain or loss achieved, controlling for relevant background characteristics including income, in a regression analysis. The coefficients from this calculation are then used to estimate the amount of income that would be required to hold life satisfaction constant after the occurrence of a particular life event. Ideally, one would perform this analysis with longitudinal panel data, so that *transitions* from one state to another can be explored, lending more confidence to the interpretation of causality and giving insight into the typical duration of subjective well-being reactions.

Using this technique, Clark and Oswald (2002) calculated that (at 1992 prices) getting married produced the same impact as an additional GBP 6 000 per month, whilst widowhood was estimated to be equivalent to losing GBP 14 000 per month. The same authors found that the impact of becoming unemployed was far greater than simply the loss of income incurred – with a monthly payment of GBP 23 000 required to offset the negative effects of unemployment. Finally, the impact of moving from "excellent" self-rated health to "fair" self-rated health was estimated as being equal to a loss of GBP 41 000 per month. To put these figures into context, the average monthly household income over the whole sample (7 500 individuals) was just under GBP 2 000. Frey, Luechinger and Stutzer (2004) have used life satisfaction data to estimate the monetary value that would be necessary to "compensate" for subjective well-being loss caused by terrorist activities in the most terrorism-prone regions of France, the United Kingdom and Ireland, as compared to the least terrorism-prone regions. Their findings indicated that between 1975 and 1998, a resident of Northern Ireland (compared to residents of Great Britain and the Republic of Ireland) experienced losses valued at 41% of total income.

Most of the work on valuation of non-market outcomes using measures of subjective well-being has focused on measures of life evaluation. However, Deaton, Fortson and Totora (2009) looked at the value of life in sub-Saharan Africa using both life evaluation measures and affect measures. They found that affect measures produced a higher value of life than that obtained using life evaluation measures.

How do the results of life satisfaction-based valuations compare to preference-based approaches? Dolan and Metcalfe (2008) note the lack of empirical research directly comparing the two, but present results from a quasi-experiment looking at the impact of an urban regeneration project in Wales. They found that whilst urban regeneration had no impact on house prices, individuals from the control group (in an adjacent area without urban regeneration) would be willing to pay on average GBP 230-245 per year for the next

three years for the public and private benefits of such a project. Meanwhile, the life satisfaction valuation method placed the impact at between GBP 6 400 and GBP 19 000 in total (over an indefinite time span).

Based on both Dolan and Metcalfe's house price data and the willingness-to-pay estimates, the GBP 10 million scheme would not look like an efficient use of resources: benefits are estimated at zero on the basis of house price differences, and at only GBP 240 000 across all households in the willingness-to-pay example. Based on the life satisfaction estimates, however, the scheme brought between GBP 6.1 million and 18.1 million in overall benefits. Such a wide range of estimates suggests a combination of approaches may be most effective in estimating value.[43]

An extension of this approach can be used in other valuation problems, including those where monetary values are not required. For example, an increasingly-used statistic in the public health field is quality-adjusted life years (QALYS). These can be used to help make decisions about which conditions and treatments to prioritise when allocating healthcare resources, particularly when investing in new healthcare technologies. According to this approach, health states are assigned a value (e.g. 0 = death; 1 = full health), and then multiplied by how long that state lasts. Dolan et al. (2009) suggest that subjective well-being data from patients could be used as a more direct way to estimate the quality-of-life improvements that result from specific health conditions and treatments. Specifically, they propose that subjective well-being could be assessed, alongside health and other important subjective well-being determinants, before and during various stages in a treatment.

Subjective well-being data can add value over other preference-based health valuation methods[44] – which involve asking either the public or patients to imagine hypothetical health-related scenarios – because both public and patient preferences can be at odds with the level of suffering actually reported by patients with those conditions (Dolan and Kahneman, 2008). For example, Dolan et al. report that in hypothetical time trade-off scenarios, individuals drawn from the general population estimate that the impact of moderate pain would be worse than the impact of moderate depression. However, when one examines both affect and life evaluation data from real patients suffering from each of these conditions, this ordering is reversed. When two subjective statements conflict, it is difficult to know which one is the "right" one; however, it seems that subjective well-being data from patients might make a useful contribution to the overall evidence base on which valuations are made.

Direction of change: Loss aversion

Evidence also suggests that particular attention should be paid to the direction of change in the variable being valued. In stated preference methods, respondents often exhibit *loss aversion* – where the negative psychological impact of a loss is expected to be greater than the corresponding positive impact of a gain in the same good (Sugden, 2005; Guria et al., 2005). Valuations based on subjective well-being data can offer critical insights into whether loss aversion reflects a *true* imbalance in the well-being derived from losses versus gains, or whether it is a result of some of the biases in decision-making noted earlier (Box 4.6). For the time being, it may be helpful to estimate valuations for losses and gains separately where possible, because the well-being impact of, for example, withdrawing a particular policy initiative may not be equivalent to the initial well-being impact of introducing it in the first place.

Using subjective well-being to adjust preference data

Where stated and revealed preference data are used in CBA, Layard, Mayraz and Nickell (2008) and Powdthavee (2010) propose that subjective well-being data can be used to resolve the problem of how to value costs and benefits accruing to people with different incomes. Direct use of prices in CBA assumes that $1 is equally valuable to all parties concerned. However, valuation surveys tend to produce very different valuation estimates for people at different points on the income distribution – because those with higher incomes are typically willing to pay more for the non-market factor in question. This can distort results because it weighs the views of higher-income individuals more heavily than the views of lower-income individuals. Layard et al. (2008) suggest using subjective well-being data (evaluative happiness, life satisfaction, or a combination of the two) to estimate the marginal utility of income. Whilst this assumes the position that subjective well-being can be used as a measure of *utility*, a position on which not everyone agrees, it nonetheless provides a way of managing the impact of the marginal utility of income in an evidence-based manner.

Challenges in the interpretation of subjective well-being valuations

The valuation of non-market factors using subjective well-being data is still in its infancy. As the analytical methods on which valuations are based are very similar to those used to investigate the drivers of subjective well-being, all of the interpretive challenges discussed in the previous section also apply here. Rather than repeating the previous section, however, the focus here will be on the implications these issues have for how monetary valuations should be conducted and interpreted – including data requirements and co-variates to include in analyses. Four factors, in particular, bear on valuations based on subjective well-being:

- *Sensitivity* of life evaluations – and what can and cannot be valued through this approach.
- *Measurement error* in estimating regression coefficients.
- Correlations among independent variables and *the co-variates to include in regression models*.
- *Time horizons* over which analyses are conducted.

Sensitivity of life evaluations

As noted earlier in this chapter, life evaluation data are sensitive to major life events and show strong associations with a variety of other well-being outcomes. However, there are *a priori* grounds not to expect large movements in life evaluations as a result of relatively small-scale policy initiatives or non-market factors. If the life satisfaction valuation technique is used to assess drivers that operate on a much smaller scale, this can risk under-valuing non-market factors that nevertheless people do regard as important. Fujiwara and Campbell (2011) also note that life evaluations may not be sensitive to the non-use value of items such as cultural monuments – so again, scale is important, and it may be more realistic to look at the collective impact of these goods.

One subjective well-being indicator that might be more sensitive to immediate surroundings and activities – for example, small changes in environmental quality, or the availability of green space – is affect. Kahneman and Sugden (2005) propose an approach to valuation based on "experienced utility". This is estimated from short-term affect data collected through the day reconstruction method, which includes information about both activities and locations, as well as the affective states accompanying those activities. The

effects of different activities or locations on positive and negative affect can then be reconstructed. Kahneman and Sugden do not propose a specific method for linking experienced utility and money – and further work is needed to address this – but they note that the life satisfaction valuation approach could be adapted for affect-based valuations.

One further limitation is that the valuation technique based on subjective well-being is retrospective, i.e. it cannot be used to project the *potential* impact of something that does not yet exist – in contrast to the hypothetical scenarios on which stated preferences are based. As policy-makers using cost-benefit analysis are frequently interested in assessing the potential effects of a policy that has not yet been put in place, analyses will often need to draw on examples of policy initiatives in other communities – where the generalisability of results to the population of interest may come with caveats.[45]

Measurement error in estimating regression coefficients

Monetary valuations obtained used subjective well-being data are typically based on regression coefficients, and thus require a high degree of precision in estimating those coefficients. Measurement error among the set of drivers (independent variables) examined in the course of valuations can be especially problematic. Of particular concern is the measurement error in self-reported income – which risks reducing the income coefficient, leading to higher valuations of non-market factors. For example, Powdthavee (2009) found an increased income coefficient (producing lower valuations for non-market factors) where objective income information was obtained by interviewers through examination of payslips.

Dolan and Metcalfe (2008) and Powdthavee (2010) also report using *instrumental variables*[46] in subjective well-being valuations to obtain better estimates of the exogenous effect of income on life evaluations. Fujiwara and Campbell (2011) note that instrumenting for income typically increases the estimate of the income coefficient, thus producing lower overall valuation estimates for non-market factors. For example, in Dolan and Metcalfe's analysis, this correction brings estimates of the value of urban regeneration down from GBP 19 000 to around GBP 7 000. Powdthavee reports that this technique lowers the valuations of marriage from around GBP 200 000 to GBP 3 500 per annum. In practice, however, it is very difficult to identify appropriate instrumental variables for the purposes of valuations, as most of the key variables strongly associated with income also tend to be associated with life satisfaction.

Measurement error in non-market factors could also reduce coefficients attached to the variable in question, leading to under-valuation. Conversely, if measurement error in non-market factors is positively correlated with measurement error in the life evaluations (for example, due to shared method variance or response biases), this could inflate rather than depress their coefficients, leading to over-valuation, unless the income variable was similarly affected. Again, instrumental variables could be of particular use in separating out causal effects from correlated errors.

The large impact that measurement error in independent variables can have on valuations means that, particularly when small or non-representative samples are used in regressions, it will be essential to check the coefficients obtained for income and other variables in the model (and especially for the non-market factor in question) to ensure that they fall within the range that might be expected, based on larger and more representative samples – and preferably those utilising high-quality panel data (Box 4.7). Further work on potential instrumental variables for use in valuations will be important for future development of the technique.

Box 4.7. **The range of income estimates observed in the life satisfaction literature**

The method used to estimate the value of non-market factors on the basis of life satisfaction data is very sensitive to the coefficient estimated for income. When interpreting the results of such valuations it is therefore helpful to consider the range of coefficients for income identified in the wider literature.

A very large number of authors have examined the role of income in life evaluations and the issue of whether increases in a country's average income over time are associated with increases in a country's subjective well-being (e.g. Easterlin, 1974, 1995; 2005; Hagerty and Veenhoven, 2003; Sacks, Stevenson and Wolfers, 2010). In practice, the effects of income are highly complex, and can vary both between countries and within different population sub-groups. Some authors report a consistent finding that income plays a more important role in developing and transition countries, and a less important role in more affluent societies (Bjørnskov, 2010; Clark, Frijters and Shields, 2008; Sacks, Stevenson and Wolfers, 2010), whereas others report a similar magnitude effect for income across all countries (Deaton, 2008; Helliwell, 2008; Helliwell and Barrington-Leigh, 2010). Estimates for income coefficients are also critically sensitive to other variables included in the regression model. Clark, Frijters and Shields (2008), Sacks, Stevenson and Wolfers (2010), and Fujiwara and Campbell (2011) provide a more extensive overview of several other important issues – including those associated with reverse causation, individual effects and the importance of relative income (i.e. an individual's income in comparison to a given reference group). A final issue is the problem of income non-response rates, which are rarely reported but could also affect coefficients estimated for income.*

Although the results from any one study should be interpreted with caution, research described below illustrates how estimates for the effect of income can vary when different background characteristics and life circumstances are controlled. In all cases, a 0-10 life evaluation measure and log-transformed income data are used.

- Sacks, Stevenson and Wolfers (2010) report results from several large life evaluation data sets, together spanning 140 countries. In cross-sectional data (pooled across all countries) and controlling only for country fixed effects, regression coefficients for log household income on Cantril Ladder life evaluations range from 0.22 to 0.28. Results remain similar when controlling for age and sex, while adjusting for the effects of permanent income or instrumenting income increases coefficients to between 0.26 and 0.5. The authors conclude that at the within-country level, the coefficient for the permanent effect of income lies somewhere between 0.3 and 0.5. They also suggest similar magnitude effects at the between-country level and for changes in income over time.
- Boarini et al. (2012) use two waves of Gallup World Poll data (2009 and 2010) to examine the determinants of Cantril Ladder life evaluations among 34 OECD countries. Pooled across countries, the coefficient for log household income is estimated at 0.18 when only key background characteristics are controlled. Controlling for a variety of other individual-level well-being outcomes (health problems, social connections, environmental quality, personal security, having enough money for food), the coefficient reduces to 0.13. When regressions for different population sub-groups are examined, the coefficient for log income is very similar for men and women (around 0.15), but much larger for those of working age (around 0.18) in comparison to the youth and the elderly (around 0.10). Drawing on a much larger number of countries involved in the Gallup World Poll (125 in total), Helliwell, Barrington-Leigh, Harris and Huang (2010) found coefficients of around 0.4 for log household income. Their analyses controlled for a wide range of variables, including demographics, social connections, religion, perceived corruption, charitable giving (time and money) and food inadequacy (not enough money for food), as well as GDP per capita and a national-level measure of food inadequacy. The food inadequacy measure was defined net of its strong and significant correlation with household income – and this raised the estimated coefficient on household income.

Box 4.7. **The range of income estimates observed in the life satisfaction literature** *(cont.)*

- Frijters, Haisken-DeNew and Shields (2004) looked at the effect of the large increase in real household income in East Germany on life satisfaction following the fall of the Berlin Wall in 1989. In the 10-year period between 1991 and 2001, the authors estimated that around 35-40% of the observed increase in average life satisfaction was attributable to the large (over 40%) increase in real household incomes during this time period, with a one-unit increase in log income corresponding to around a 0.5 unit increase in life satisfaction for both men and women.
- Dolan and Metcalfe (2008) examined the life satisfaction impact of an urban regeneration project in a quasi-experimental design, which involved comparing two different communities in Wales. Background variables such as gender, age, relationship status and employment status were controlled in analyses. Whole sample analyses failed to find a significant effect of log household income on life satisfaction. With analyses restricted to individuals of working age, coefficients for household income were observed in the range of 0.65 to 0.93. These authors also note that in additional analyses, controlling for measures of social capital reduced the income coefficient by a non-trivial amount.

* Although hard data is rarely reported, it appears that a relatively high proportion of individuals refuse to answer questions about their income in non-official and telephone-based surveys. For example, Smith (2013) estimated that income non-response rates ranged from under 5% to just over 35% for countries involved in the 2008 European Values Survey, with a similar variation in the World Values Survey. Gasparini and Gluzmann (2009) found non-response rates in the 2006 Gallup World Poll among Latin American and Caribbean countries to range from 2% (in Ecuador) to 39% (in Trinidad and Tobago). The potentially non-random nature of income non-responses (Riphahn and Serfling, 2005; Gasparini and Gluzmann, 2009; Zweimüller, 1992) and the various techniques deployed to manage them (which range from dropping all observations from the analysis, to imputation methods) can potentially impact on estimated coefficients. Further research and reporting on this phenomenon is needed.

Co-variates to include in the regression model

As noted earlier, any attempts to compare coefficients obtained through regression analyses need to consider the possible impact of correlations among independent variables. When using life satisfaction data to value non-market factors, Fujiwara and Campbell (2011) recommend that measures for *all* known determinants of well-being should be included in the model. Although they note the lack of consensus in this area, they list key determinants from the literature as: income, age, gender, marital status, educational status, employment status, health status, social relations, religious affiliation, housing, environmental conditions, local crime levels, number of children and other dependents (including caring duties), geographic region, personality traits (such as extroversion) and the non-market factor being valued.

The same authors also note that for policy purposes, there may be some indirect effects that need to be controlled in valuation regressions to fully estimate the impact that a *marginal change* in a non-market factor may have on subjective well-being. They take the example of pollution, noting that although pollution is expected to have negative effects on subjective well-being, individuals may be partially compensated for those effects through lower house prices and reduced commuting times. These offset the overall impact of pollution on subjective well-being and may cause the *true* value of a marginal reduction in pollution to be underestimated. Frey, Luechinger and Stutzer (2004) note the same difficulty in estimating the impact of living in terrorism-prone areas, where higher wages and lower rents potentially compensate individuals – and these authors conclude that all potential channels of compensation need to be controlled for.

The difficulty in attempting to control for all possible drivers and indirect effects is that this may crowd out the variables of interest. For example, including controls that also co-vary with income in the regression equation may shrink the coefficient for income, thus shrinking the increase in life satisfaction brought about by each additional per cent increase in income. Underestimating the impact of income can therefore risk over-valuing the impact of non-market factors. However, the same under-valuation risk is also present for the non-market factors. For example, if the effects of air pollution on life evaluations are mediated by respiratory health conditions, the coefficient for a measure of air pollution is likely to be substantially reduced if a measure of respiratory health conditions is included in the model. This would lead to a lower valuation of air pollution than if the health variable were excluded from the model. The choice to include or exclude other variables in the regression therefore depends on the assumed causal pathways – and these must be clearly described when conducting valuations. Mediational analyses therefore need to play an essential role in preparing models, to better understand how predictors interact with one another. In further developing this valuation technique, it is also important to establish a better overall consensus regarding which explanatory variables should be included in (and excluded from) a valuation regression, and under what circumstances. As noted previously, it may be helpful to report results as a range of values, derived from various different models with and without the presence of certain control variables.

Time horizons

Whereas more traditional approaches to valuation enable time horizons to be specified if necessary (i.e. a period of time over which respondents would be willing to pay, or willing to accept, a particular non-market good), the subjective well-being valuation approach does not come with a fixed time-frame. Dolan and Metcalfe (2008) note that current valuation approaches fail to address the issue of how long changes in subjective well-being are thought to last as a result of the effects of a particular non-market factor.[47]

Time horizons are also important for being able to adequately *detect* and value the impact of events, interventions or non-market factors. Consideration of the mechanisms through which subjective well-being impacts are realised, and the time frames over which those mechanisms operate, is thus important in selecting data sets for valuation purposes and in interpreting the findings (Frey, Luechinger and Stutzer, 2004). The potential for partial adaptation to life events over time (see Section 2) also has implications for valuations. For example, the subjective well-being impact of widowhood ten years after the event is likely to be different to the subjective well-being impact just one year after the event. Thus, for the valuation of exogenous events, it is important to specify, *a priori*, the time frame(s) of interest and to collect, examine and interpret the data accordingly. In particular, scope exists for measuring effects at given points in time following a change, as well as for establishing longer-term averages or cumulants.

Combining non-market and market prices in cost-benefit analysis

The ultimate goal of assigning monetary values to non-market goods and services is to enable them to be examined in cost-benefit analyses alongside goods and services with market prices. Because the subjective well-being valuation technique is still in its infancy, and tends to produce wide-ranging estimates of value, the UK government Treasury's current position (as articulated in the 2011 update to *The Green Book: Appraisal and Evaluation in Central Government*) is that whilst valuations based on life satisfaction might be a useful

way to quantify the *relative* value of two or more non-market goods, they are not yet robust enough to allow comparisons with market prices. An alternative to this view would be to generate and test a series of cost-benefit analysis models – starting with a benchmark model based on market prices and other estimations widely regarded as robust, and extending this with valuations based on preference-based approaches on the one hand, and subjective well-being on the other.

One example of this multi-stage approach is the work of Gyarmati et al. (2008), who undertook a comprehensive quasi-experimental evaluation of the Community Employment Innovation Project in Canada in order to estimate the overall costs and benefits of the project to individuals, communities and governments. The project was designed to evaluate a long-term active re-employment strategy for unemployed individuals who volunteered to work on locally-developed community projects in return for wages (as an alternative to receiving state-funded income transfers). A benchmark cost-benefit analysis model was based on administrative costs, participant earnings and the market prices of fringe benefits, taxes, transfer payments and premiums, as well as market-based estimations of the value of volunteering. An extended model was then developed that included a valuation of foregone leisure (based on 20% of earnings), as well as a valuation based on subjective well-being (based on Helliwell and Huang, 2005) of the social networks that respondents built as a result of their work placements, and the reduction in perceived hardship experienced. In the benchmark model, each dollar in net cost to government was estimated to produce between $1.02 and $1.39 in net benefits to society.[48] In the extended model, each dollar in net cost to government was estimated to bring between $1.21 and $1.61 of net benefits. Gyarmati et al. point out that one dollar direct cash transfer (one alternative to the employment project) has meanwhile been estimated to deliver only $0.85 in net benefits to the intended recipient.

Conclusions

Fujiwara and Campbell (2011) summarise the advantages of the life satisfaction approach to valuation as follows: i) the cost and time-effectiveness of the data collection; ii) the ability to use statistics drawn from very large and representative samples (in contrast to stated preference techniques, which require a separate data collection exercise); iii) the possible application to a whole variety of life events and circumstances; iv) the presence of fewer biases and less strategic behaviour on the part of respondents; and v) the fact that the data do not rest on assumptions about market structure. Conversely, some of the disadvantages highlighted by these authors include: i) difficulties in estimating the marginal utility of income, including the effects of relative as compared to absolute income, as well as the indirect effects of income and variables that operate counter to the effects of income; and ii) difficulties in estimating the marginal utility of the non-market factor, including indirect effects and the consumption of complementary goods alongside the non-market factor. All of the main approaches to monetary valuation of non-market factors (revealed preference, stated preference and subjective well-being-based estimates) are associated with methodological shortcomings, but the nature of these shortcomings is different in each case. Using several different methods provides more information than relying on a single approach, and using subjective well-being data offers a relatively low-cost option that avoids some of the biases connected to preference-based approaches. However, as the method based on subjective well-being is still in its infancy, significant methodological and interpretive questions remain, and it should therefore be regarded as a complement to rather than a replacement for existing methods.

Notes

1. *http://hdr.undp.org/en/statistics/hdi/*.
2. It is important to note that these findings are based on worldwide visitors to the OECD's *Your Better Life Index* website, *http://oecdbetterlifeindex.org/*, a sample of individuals known to be non-representative and non-random, and as such they should be interpreted with care.
3. *Frame-of-reference effects* refer to differences in the way respondents formulate their answers to survey questions, based on their own life experiences, as well as their knowledge about the experiences of others – including both those they consider as within their "comparison group" and those outside it.
4. *Adaptation* refers to psychological processes that may either restore or partially repair subjective well-being, and particularly affective experiences, in the face of some types of adversity. People may also show adaptation to positive life events over time (whereby the initial subjective well-being boost delivered by a positive change in life circumstances, such as marriage, reduces over time).
5. Much of the critique surrounding the use of subjective well-being for public policy centres around a view that increasing positive emotions is not an appropriate goal for governments (e.g. Booth et al., 2012; McCloskey, 2012). Although this view potentially underestimates the health and well-being implications of emotional experiences (described above), and fails to distinguish between the usefulness of *monitoring* positive emotions, versus making them *primary objectives* of government policy, it is nonetheless further grounds to avoid describing subjective well-being data solely in terms of "happiness".
6. Ordinal data are those measured on scales where the intervals between scale points are not assumed to be equal, but there is an underlying sequence or rank order. For example, we assume that a 5 is lower than a 6 and a 6 is lower than a 7, but we do not assume that the distance between 5 and 6 is equivalent to the distance between 6 and 7. Linear regression relies on continuous variables, where cardinality is assumed, i.e. where the size of the number on a scale is expected to have a direct linear relationship with the amount of the variable in question. Tabachnick and Fidell (2001), however, note that in the social sciences, it is common practice to treat ordinal variables as continuous, particularly where the number of categories is large – e.g. seven or more – and the data meet other assumptions of the analysis.
7. Data users are likely to want to know, for example, *how* good or bad a person's experience was, not just on which side of a cut-off they fall. This is less of a problem for the reporting of headline national aggregate figures, but becomes particularly relevant when comparing responses between groups. It can be ameliorated to some extent by banding responses into several categories rather than selecting just one cut-off point.
8. For example, a country with universally low levels of subjective well-being would have few individuals falling below the relative poverty line, thus masking the extent of difficulties faced.
9. For example, the thresholds associated with clinically-significant mental health outcomes may be very different to the thresholds associated with different educational or income levels.
10. There are some who disagree with this, arguing that cardinal interpretations of subjective well-being are possible – e.g. Ng (1997).
11. It is unclear, for example, what it really means to say that the bottom 10% of the population achieves only 1% of the total subjective well-being. This can be contrasted with income, where it is easier to understand the practical implications of the bottom 10% earning just 1% of the total income across a population.
12. Although Helliwell, Barrington-Leigh, Harris and Huang (2010) took the simple mean average of the Cantril Ladder and a single-item life satisfaction measure and found this was more closely correlated with predictors of subjective well-being (such as demographics, income and a set of social indicators) than either measure on its own.
13. The UK's ONS have also proposed a single-item eudaimonia question, for high-level monitoring purposes: "Overall, to what extent do you feel the things you do in your life are worthwhile?" (ONS, 2011b).
14. In the case of more detailed analyses, such as group comparisons or investigation of the drivers of subjective well-being, separate estimates of the different sub-components of subjective well-being will be preferred due to the risk of information loss when summing across sub-components.

15. For example, if the threshold on a 0-10 scale is set at 7, movements across that threshold will be very salient, but large-scale movements from 8 to 9, or from 2 to 5, may go undetected.

16. Represented as a percentage of the population reporting higher positive than negative affect; OECD calculations based on figures from the 2010 Gallup World Poll.

17. I.e. a change over a one, five or ten-year period. As discussed earlier, short-term fluctuations of this magnitude can also be detected, but may not represent meaningful societal shifts in overall levels of well-being (e.g. Deaton, 2012).

18. For example, the *World Happiness Report* (Figure 2.3) lists 63 countries where the mean average life evaluation between 2005 and 2011 (measured on a 0-10 Cantril Ladder scale) is lower than the scale midpoint, 5. These include India, China, Iraq, Afghanistan and Syria; and particularly low-scoring countries, with scores below 4.0, include Congo, Tanzania, Haiti, Comoros, Burundi, Sierra Leone, the Central African Republic, Benin and Togo.

19. *Reverse causality* in this context refers to when subjective well-being drives the independent variable, rather than vice versa. For example, in a cross-sectional analysis, a significant association could be observed between income (the independent variable) and subjective well-being (the dependent variable), but this could be because subjective well-being drives income (rather than vice versa). *Two-way causality* is where there are reciprocal and causal relationships between two variables in both directions – i.e. income drives subjective well-being, but subjective well-being also drives income. *Endogeneity* refers to a situation where there is a correlation between an independent variable and the error term in a regression model. This can be due to measurement error, omitted variables, sample selection errors, and/or reverse or two-way causality.

20. To quote from Kahneman and Riis (2005): "… consider the Americans and the French. The distributions of life satisfaction in the US and France differ by about half a standard deviation. For comparison, this is also the difference of life satisfaction between the employed and the unemployed in the US, and it is almost as large as the difference between US respondents whose household income exceeds USD 75 000 and others whose household income is between USD 10 000 and USD 20 000 (in 1995)… Is it possible to infer from the large differences in evaluated well-being that experienced well-being is also much lower in France than in the USA? We doubt it, because the sheer size of the difference seems implausible" (p. 297).

21. Such as tendencies to use either extreme or more "moderate" scale response categories, as well as the likelihood of socially desirable responding.

22. There are times when people's subjective perceptions matter, even when they don't reflect objective reality: "cultural differences may in some cases be relevant to policy and in some cases irrelevant. For example, people's satisfaction with leisure opportunities might be relevant to policy deliberations, regardless of the objective conditions" (Diener, Inglehart, and Tay, 2012, p. 20). Another classic example of this is perceptions of regulatory burden, which can influence important business decisions and behaviour regardless of their accuracy (OECD, 2012). If perceptions of regulatory burden are misplaced, activity should be focused on better informing businesses and the wider public. Few would argue that the correct response would be to simply adjust the perceptions of regulatory burden so that they fit the pattern observed among more objective measures.

23. This Latin American effect has been explored in depth by Graham and Lora (2009).

24. See Chapter 2 for a full account of response styles.

25. As Senik (2011) puts it, "If the French evaluate the happiness of some hypothetical person in a less positive manner than the Danes, perhaps it is because they would actually feel less happy in the situation of that hypothetical person" (p. 8).

26. In practice, however, substituting even the highest adjusted figure for French natives (7.54) would only cause a very minor adjustment in country rankings overall, causing French natives (7.22) to exchange places with natives of Great Britain (7.38) only. Overall, mean average happiness ratings among natives in the 13 countries sampled range from 6.74 in the case of Portugal to 8.34 in the case of Denmark.

27. Suppose that two individuals both have the same levels of underlying positive and negative affect. But imagine that the first has a tendency towards extreme responding – so that on a 0-10 scale this individual reports 9 on positive affect and 7 on negative affect. The second individual has a tendency towards more moderate responding, thus reporting a 7 on positive affect and a 5 on negative affect. The net affect balance for both individuals will be +2. This of course assumes that response biases operate in a similar way for both positive and negative affects, which requires further examination.

28. A substantial part of the literature in this field uses *country* as a proxy for culture, inferring cultural differences on the basis of country differences. Whilst this is potentially problematic, the words country and culture are often used interchangeably in many studies on the subject.

29. Although the implications of the subjective well-being literature are often interpreted in terms of opportunities for policy *interventions*, subjective well-being has just as much potential to identify areas where existing government interventions can be redesigned or stopped altogether.

30. Double-blind conditions refer to scenarios where neither the respondent nor those implementing the intervention are aware of which treatment group a given respondent has been assigned to. Single-blind is where those implementing the intervention know which treatment group has been assigned to which respondent, but respondents are unaware.

31. Wider applicability can be challenged where there are concerns about the extent to which a given result might generalise to other situations, beyond the experimental or quasi-experimental setting.

32. For some research questions investigating international differences in subjective well-being, where the driver in question is hypothesised to operate at an aggregate country level, pooled cross-sectional time series data (i.e. international data containing repeated study waves and representative, but different, samples in each wave) may also enable some causal inferences.

33. Factor analysis is a statistical procedure that is conducted to identify distinct (relatively independent) clusters or groups of related items or variables (called factors). It is based on patterns of correlations among items or variables.

34. Because error is estimated and removed in the process of extracting the underlying factors. A "factor loading" is calculated for each item or variable, which reflects the variance it shares with the underlying factor – and all other variance is assumed to be error. When factors are used in the analysis (instead of measured variables), only this common variance is analysed, and thus measurement error is, in theory, purged from the data.

35. One interesting exception, however, is cultural biases. Although there is currently some evidence of cultural bias in direct country comparisons of mean average levels, there is currently little to suggest that cultural biases exert a problematic influence on cross-country analyses of the drivers of subjective well-being – and drivers tend to be reasonably consistent across countries (e.g. Fleche, Smith and Sorsa, 2011; Helliwell and Barrington-Leigh, 2010). The issue of the extent to which regression solutions are replicable and can be generalised from one sample or country to another is, however, an important consideration that is discussed below.

36. However, statistically, mediation and confounding are identical: both are indicated when the inclusion of the third variable in the model reduces the relationship between the independent variable and the dependent variable by a non-trivial amount.

37. An example would be a significant statistical relationship between the time I spend talking to the plant in my office and how much it grows per year – both of which are related to an (unmeasured) causal variable: how often I remember to water the plant.

38. An example here might be that a causal relationship between how often I remember to water my office plant and how much it grows is obscured by a third (unmeasured) variable: how much plant food my colleague gives it.

39. Shared method variance refers to variance that is attributed to the measurement method, rather than the constructs of interest. In the case of subjective well-being, the main concern is that if drivers are also measured through subjective self-report data, self-report biases (including retrospective recall biases, response styles, cultural bias, etc.) could inflate observed relationships between those drivers and the subjective well-being outcomes of interest.

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

41. An instrumental variable is one that has a direct association with the independent variable in question (e.g. income), but not with the outcome of interest (e.g. life evaluations).

42. The term *function* is used here to describe the overall pattern of relationships between the independent variables and the dependent variable, including the size and significance of coefficients.

43. Dolan and Metcalfe conclude that "we need much more research into the extent and the sources of the differences between these valuation methods" (p. 25), particularly given that the valuation through subjective well-being approach is still in its infancy and "literally thirty years behind that of generating monetary values from revealed and stated preferences".

44. These include the "standard gamble" and "time trade-off" methods (see Dolan and Kahneman, 2008).

45. See the previous section on interpreting the drivers of subjective well-being for a more detailed treatment of the generalisability of results.

46. An instrumental variable is one that has a direct association with the independent variable in question (e.g. income), but that is associated with the outcome of interest (e.g. life evaluations) only via the independent variable in question.

47. Based on the current state of knowledge, these authors suggest that "there would seem to be good grounds for viewing the ICs (income compensation – i.e. valuations) as a total value over a finite horizon. Clearly, the actual assumption made on how life satisfaction incorporates future expectations is crucial to the methodology of the value of the non-market good by experiences, and merits further investigation" (p. 23).

48. Two different estimates were produced because different models were estimated for participants, based on which type of welfare payments they received from the government prior to their participation in the programme.

Bibliography

Abelson, R.P. (1985), "A variance explanation paradox: When a little is a lot", *Psychological Bulletin*, Vol. 97(1), pp. 129-133.

Angelini, V., D. Cavapozzi, L. Corazzini and O. Paccagnell (2011), "Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases", *Health, Econometrics and Data Group Working Paper*, No. 11/20, University of York, available online at: *www.york.ac.uk/res/herc/hedgwp*.

Ariely, D., G. Loewenstein and D. Prelec (2003), "Coherent arbitrariness: Stable demand curves without stable preferences", *The Quarterly Journal of Economics*, Vol. 118, pp. 73-105.

Beegle, K., K. Himelein and M. Ravallion (2012), "Frame-of-reference bias in subjective welfare", *Journal of Economic Behaviour and Organization*, Vol. 81, pp. 556-570.

Bertrand, M. and S. Mullainathan (2001), "Do people mean what they say? Implications for subjective survey data", *The American Economic Review*, Vol. 92(2), pp. 67-72.

Bjørnskov, C. (2012), "Wellbeing and the size of government", in P. Booth (ed.), *… and the Pursuit of Happiness: Wellbeing and the role of Government*, London: Institute of Economic Affairs, in association with Profile Books.

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

Blanchflower, D.G. (2009), "International evidence on well-being", in *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, pp. 155-226, University of Chicago Press.

Blanchflower, D.G. and A.J. Oswald (2008), "Is well-being U-shaped over the life cycle?", *Social Science and Medicine*, Vol. 66(8), pp. 1733-1749.

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

Boarini, R., M. Comola, C. Smith, R. Manchin and F. De Keulenaer (2012), "What Makes for a Better Life? The Determinants of Subjective Well-Being in OECD Countries – Evidence from the Gallup World Poll", *OECD Statistics Working Papers*, No. 2012/03, OECD Publishing, DOI: *http://dx.doi.org/10.1787/5k9b9ltjm937-en*.

Bok, D. (2010), *The Politics of Happiness: What government can learn from the new research on well-being*, Princeton and Oxford: Princeton University Press.

Bolger, N. and A. Zuckerman (1995), "A framework for studying personality in the stress process", *Journal of Personality and Social Psychology*, Vol. 69(5), pp. 890-902.

Booth, P. (ed.) (2012), *… and the Pursuit of Happiness: Wellbeing and the role of government*, London: The Institute of Economic Affairs, available online at: *www.iea.org.uk/publications/research/and-the-pursuit-of-happiness*, last accessed 10 August 2012.

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*, San Francisco: Jossey-Bass.

Brickman, P., D. Coates and R. Janoff-Bulman (1978), "Lottery winners and accident victims: Is happiness relative?", *Journal of Personality and Social Psychology*, Vol. 36(8), pp. 917-927.

Brief, A.P., M.J. Burke, J.M. George, B.S. Robinson and J. Webster (1988), "Should negative affectivity remain an unmeasured variable in the study of job stress?", *Journal of Applied Psychology*, Vol. 73(2), pp. 193-198.

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.

Carnevale, P.J.D. and A.M. Isen (1986), "The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiations", *Organizational Behavior and Human Decision Processes*, Vol. 37(1), pp. 1-13.

Chapple et al. (2010), "Subjective well-being and social policy", *European Commission Working Paper*, published in the *Social Europe* series, under the Directorate-General for Employment, Social Affairs and Inclusion, available online at: *http://ec.europa.eu/social/main.jsp?langId=en&catId=22*, last accessed 10 August 2012.

Clark, A.E. (2003), "Unemployment as a social norm: Psychological evidence from panel data", *Journal of Labor Economics*, Vol. 21, No. 2, pp. 323-351.

Clark, A.E., E. Diener, Y. Georgellis and R.E. Lucas (2008), "Lags and leads in life satisfaction: A test of the baseline hypothesis", *The Economic Journal*, Vol. 118, pp. 22-243.

Clark, A.E., P. Frijters and M. Shields (2008), "Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles", *Journal of Economic Literature*, Vol. 46(1), pp. 95-144.

Clark, A.E. and A.J. Oswald (2002), "A simple statistical method for measuring how life events affect happiness", *International Journal of Epidemiology*, Vol. 31, pp. 1139-1144.

CLES Consulting and the New Economics Foundation (2011), "Big Lottery Fund national well-being evaluation", progress report and interim findings, prepared by CLES Consulting and presented to the Big Lottery Fund, February, available online at: *www.biglotteryfund.org.uk/evaluation_well-being.htm?fromsearch=-uk*, last accessed 9 August 2012.

Cohen, S., W.J. Doyle, R.B. Turner, C.M. Alper and D.P. Skoner (2003), "Emotional style and susceptibility to the common cold", *Psychosomatic Medicine*, Vol. 65(4), pp. 652-657.

Cohen, S. and S.D. Pressman (2006), "Positive affect and health", *Current Directions in Psychological Science*, Vol. 15, pp. 122-125.

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

Cummins, R.A. (2003b), "Subjective well-being from rich and poor", Social Indicators Research Series, Vol. 15, Part 3, pp. 137-156.

Cummins, R.A. (2001), "The subjective well-being of people caring for a family member with a severe disability at home: a review", *Journal of Intellectual and Developmental Disability*, Vol. 26(1), pp. 83-100.

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.

Danner, D., D. Snowdon and W. Friesen (2001), "Positive emotions in early life and longevity: Findings from the Nun Study", *Journal of Personality and Social Psychology*, Vol. 80, pp. 804-813.

Deaton, A. (2012), "The financial crisis and the well-being of Americans; 2011 OEP Hicks Lecture", *Oxford Economic Papers*, Vol. 64, pp. 1-26.

Deaton, A. (2010), "Income, aging, health and well-being around the world: Evidence from the Gallup World Poll", in D.A. Wise (ed.), *Research Findings in the Economics of Aging*, Chicago: University of Chicago Press, pp. 235-263.

Deaton, A. (2008), "Income, aging, health and well-being around the world: Evidence from the Gallup World Poll", *NBER Working Paper*, No. 13317, National Bureau of Economic Research.

Deaton, A., J. Fortson and R. Totora (2009), "Life (evaluation), HIV/AIDS, and death in Africa", *NBER Working Paper*, No. 14637, National Bureau of Economic Research.

Di Tella, R. and R. MacCulloch (2010), "Happiness adaption to income beyond "basic needs"", in E. Diener, J. Helliwell and D. Kahneman (eds.), *International Differences in Well-being*, New York: Oxford University Press.

Di Tella, R. and R. MacCulloch (2008), "Gross national happiness as an answer to the Easterlin Paradox?", *Journal of Development Economics*, Vol. 86, pp. 22-42.

Di Tella, R., R. MacCulloch and A.J. Oswald (2003), "The Macroeconomics of Happiness", *The Review of Economics and Statistics*, Vol. 85(4), pp. 809-827.

Di Tella, R., R. MacCulloch and A.J. Oswald (2001), "Preferences over inflation and unemployment: Evidence from surveys of happiness", *The American Economic Review*, Vol. 91(1), pp. 335-341.

Diener, E., M. Diener and C. Diener (1995), "Factors predicting the subjective well-being of nations", *Journal of Personality and Social Psychology*, Vol. 69, pp. 851-864.

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*, Social Indicators Research Series, Vol. 39, Dordrecht: Springer, pp. 247-266.

Diener, E., R.E. Lucas and C. Napa Scollon (2006), "Beyond the hedonic treadmill: Revising the adaptation theory of well-being.", *American Psychologist*, Vol. 61(4), pp. 305-314.

Diener, E., R.E. Lucas, U. Schimmack and J. Helliwell (2009), *Well-being for public policy*, New York: Oxford University Press.

Diener, E., C. Nickerson, R.E. Lucas and E. Sandvik (2002), "Dispositional affect and job outcomes", *Social Indicators Research*, Vol. 59(3), pp. 229-259.

Diener, E., C.K.N. 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. and M.E.P. Seligman (2004), "Beyond money: Towards an economy of well-being", *Psychological Science in the Public Interest*, Vol. 5(1), pp. 1-31.

Diener, E. and W. Tov (2012), "National Accounts of Well-Being", in K.C. Land, A.C. Michalos and M.J. Sirgy (eds.), *Handbook of Social Indicators and Quality of Life Research*, Springer: Dordrecht.

Dolan, P. (2009), "NICE should value real experiences over hypothetical opinions", letter to *Nature*, published Vol. 462(5), p. 35.

Dolan, P. and D. Kahneman (2008), "Interpretations of utility and their implications for the valuation of health", *The Economic Journal*, Vol. 118, pp. 215-234.

Dolan, P., H. Lee, D. King and R. Metcalfe (2009), "How does NICE value health?", *British Medical Journal*, Vol. 229, pp. 371-373.

Dolan, P. and R. Metcalfe (2011), "Comparing measures of subjective well-being and views about the role they should play in policy", *UK Office for National Statistics Paper*, July.

Dolan, P. and R. Metcalfe (2008), "Comparing willingness-to-pay and subjective well-being in the context of non-market goods", *CEP Discussion Paper*, No. 890, London: Centre for Economic Performance, London School of Economics.

Dolan, P., T. Peasgood and M. White (2007), "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being", *Journal of Economic Psychology*, Vol. 29, pp. 94-122.

Dolan, P. and M.P. White (2007), "How can measures of subjective well-being be used to inform policy?", *Perspectives on Psychological Science*, Vol. 2(1), pp. 71-85.

Easterlin, R.A. (2005), "Feeding the illusion of growth and happiness: A reply to Hagerty and Veenhoven", *Social Indicators Research*, Vol. 74(3), pp. 429-433.

Easterlin, R.A. (1995), "Will raising the incomes of all increase the happiness of all?", *Journal of Economic Behaviour and Organisation*, Vol. 27(1), pp. 35-48.

Easterlin, R.A. (1974), "Does Economic Growth Improve the Human Lot? Some Empirical Evidence", in David, P.A. and M.W. Reder, *Nations and Households in Economic Growth: Essays in Honour of Moses Abramovitz*, New York, Academic Press Inc, pp. 89-125.

Easterlin, R.A., R. Morgan, M. Switek and F. Wang (2012), "China's life satisfaction, 1990-2010", *Proceedings of the National Academy of Sciences*, early edition 2012 (contributed 06 April 2012).

Estrada, C.A., A.M. Isen and M.J. Young (1997), "Positive affect facilitates integration of information and decreases anchoring among Physicians", *Organizational Behavior and Human Decision Processes*, Vol. 72(1), pp. 117-135.

Eurostat (2009), *Sustainable Development in the European Union: 2009 monitoring report of the EU sustainabledevelopment strategy*, Eurostat Statistical Books, available online at: *http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-78-09-865/EN/KS-78-09-865-EN.PDF*.

Ferrer-i-Carbonell, A. and P. Frijters (2004), "How important is methodology for estimates of the determinants of happiness?", *The Economic Journal*, Vol. 114, pp. 641-659.

Fitoussi, J.P. and J. Stiglitz (2011), "On the measurement of social progress and well being: Some further thoughts", paper presented at the 16th World Congress of the International Economic Association, Beijing, 4-8 July.

Fleche, S., C. Smith and P. Sorsa (2011), "Exploring Determinants of Subjective Wellbeing in OECD Countries: Evidence from the World Value Survey", *OECD Economics Department Working Papers*, No. 921, OECD Publishing, DOI: *http://dx.doi.org/10.1787/5kg0k6zlcm5k-en*.

Frey, B., C. Benesch and A. Stutzer (2007), "Does watching TV make us happy?", *Journal of Economic Psychology*, Vol. 28, pp. 283-313.

Frey, B.S., S. Luechinger and A. Stutzer (2004), "Valuing public goods: The life satisfaction approach", *CESifo Working Paper*, No. 1158, Zurich: University of Zurich Institute for Empirical Research in Economics.

Frey, B.S. and A. Stutzer (2002), "What can economists learn from happiness research?", *Journal of Economic Literature*, Vol. 40(2), pp. 402-435.

Frey, B.S. and A. Stutzer (2000), "Happiness, economy and institutions", *The Economic Journal*, Vol. 110(446), pp. 918-938.

Frijters, P., J.P. Haisken-DeNew and M.A. Shields (2004), "Money does matter! Evidence from increasing real income and life satisfaction in East Germany following reunification", *American Economic Review*, Vol. 94, pp. 730-40.

Fujiwara, D. and R. Campbell (2011), *Valuation techniques for social cost-benefit analysis: Stated preference, revealed preference and subjective well-being approaches, a discussion of the current issues*, United Kingdom: HM Treasury and Department for Work and Pensions, July.

Gasparini, L. and P. Gluzmann (2009), "Estimating income poverty and inequality from the Gallup World Poll: The case of Latin America and the Caribbean", *ECINEQ Working Paper*, 2009-151, published December 2009, Palma de Mallorca, ECINEQ Society for the Study of Economic Inequality, available online at: *www.ecineq.org/milano/wp/ecineq2009-151.pdf*, last accessed 14 August 2012.

Gilbert, D.T., M.D. Lieberman, C.K. Morewedge and T.D. Wilson (2004), "The peculiar longevity of things not so bad", *Psychological Science*, Vol. 15(4), pp. 14-19.

Gilbert, D.T. and T.D. Wilson (2000), "Miswanting: Some problems in the forecasting of future affective states", in J.P. Forgas (ed.) (2000), *Feeling and thinking: The role of affect in social cognition*, Studies in emotion and social interaction, second series, pp. 178-197, New York, NY, US: Cambridge University Press.

Godefroy, P. (2011), "Life-satisfaction: French people give themselves an average score of 7 out of 10", *INSEE, France, portrait social – édition 2011*, available online at: *www.insee.fr/en/publications-et-services/dossiers_web/stiglitz/VE4-Anglais.pdf*.

Graham, C. and E. Lora (eds.) (2009), *Paradox and Perception: Measuring Quality of Life in Latin America*, Washington, DC, The Brookings Institution Press.

Gross, E.F., J. Juvonen and S.L. Gable (2002), "Internet use and well-being in adolescence", *Journal of Social Issues*, Vol. 58(1), pp. 75-90.

Gruber, J. and S. Mullainathan (2002), "Do cigarette taxes make smokers happier?", *NBER Working Paper*, No. 8872, Cambridge, Mass.: National Bureau of Economic Research.

Guria, J., J. Leung, J. Jones-Lee and G. Loomes (2005), "The willingness to accept value of statistical life relative to the willingness-to-pay value: Evidence and policy implications", *Environmental and Resource Economics*, Vol. 32, pp. 113-127.

Gyarmati, D., S. de Radd, B. Palameta, C. Nicholson and T. Shek-Wai Hui (2008), "Encouraging work and supporting communities: Final results of the Community Employment Innovation Project", Ottawa: Social Research and Demonstration Corporation, available online at: *www.srdc.org/uploads/CEIP_finalrpt_ENG.pdf*, last accessed 10 August 2012.

Hagerty, M.R. and R. Veenhoven (2003), "Wealth and happiness revisited: Growing wealth of nations *does* go with greater happiness", *Social Indicators Research*, Vol. 64, pp. 1-27.

Halpern, D. (2010), *The Hidden Wealth of Nations*, Cambridge: Polity Press.

Helliwell, J.F. (2008), "Life satisfaction and quality of development", *NBER Working Papers*, No. 14507, National Bureau of Economic Research.

Helliwell, J.F. and C.P. Barrington-Leigh (2010), "Measuring and understanding subjective well-being", *NBER Working Paper*, No. 15887, National Bureau of Economic Research.

Helliwell, J.F., C.P. Barrington-Leigh, A. Harris and H. Huang (2010), "International evidence on the social context of well-being", in E. Diener, J.F. Helliwell and D. Kahneman (eds.), *International Differences in Well-Being*, New York: Oxford University Press.

Helliwell, J.F. and H. Huang (2005), "How's the job? Well-being and social capital in the workplace", Cambridge: National Bureau of Economic Research, available online at: *http://faculty.arts.ubc.ca/jhelliwell//papers/NBERw11759.pdf*, last accessed 10 August 2012.

Helliwell, J.F., R. Layard and J. Sachs (eds.) (2012), *World Happiness Report*, Earth Institute, New York, Columbia University, available online at: *www.earthinstitute.columbia.edu/sitefiles/file/Sachs%20Writing/2012/World%20Happiness%20Report.pdf*.

HM Treasury (2011), *The Green Book: Appraisal and Evaluation in Central Government*, United Kingdom.

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.

Huxley, A. (1932; reprinted 2004), *Brave New World*, London: Vintage Classics.

Inglehart, R., R. Foa, C. Peterson and C. Welzel (2008), "Development, freedom and rising happiness: A global perspective 1981-2007", *Perspectives on Psychological Science*, Vol. 3, pp. 264-285.

International Wellbeing Group (2006), *Personal Wellbeing Index*, 4th edition, Melbourne: Australian Centre on Quality of Life, Deakin University, available online at: *www.deakin.edu.au/research/acqol/instruments/wellbeing_index.htm*.

ISTAT (2011), *The Importance of Well-Being Dimensions for Citizens*, web article published 4 November, available online at: *www.istat.it/en/archive/44236*, accessed 09 August 2012.

Jürges, H. (2007), "True health vs response styles: Exploring cross-country differences in self-reported health", *Health Economics*, Vol. 16, pp. 163-178.

Kahneman, D. (2010), "The riddle of experience versus memory", lecture delivered February 2010 as part of the *TED Talks* series, available online at: *www.ted.com/talks/daniel_kahneman_the_riddle_of_experience_vs_memory.html*.

Kahneman, D. (1999), "Objective happiness", in D. Kahneman, E. Diener and N. Schwarz (eds.), *Well-being: Foundations of hedonic psychology*, pp. 3-25, New York: Russell Sage Foundation Press.

Kahneman, D. and A. Deaton (2010), "High income improves life evaluation but not emotional well-being", *Proceedings of the National Academy of Sciences*, Vol. 207(38), pp. 16489-16493.

Kahneman, D. and A.B. Krueger (2006), "Developments in the measurement of subjective well-being", *The Journal of Economic Perspectives*, Vol. 20(1), pp. 3-24.

Kahneman, D., A.B. Kreuger, D. Schkade, N. Schwartz and A. Stone (2004), "Towards National Well-Being Accounts", *The American Economic Review*, Vol. 94(2), pp. 429-434.

Kahneman, D. and J. Riis (2005), "Living and thinking about it: Two perspectives on life", in F.A. Huppert, B. Kaverne and N. Baylis (eds.), The Science of Well-Being, London: Oxford University Press, pp. 285-304.

Kahneman, D. and R. Sugden (2005), "Experienced utility as a standard of policy evaluation", *Environmental and Resource Economics*, Vol. 32, pp. 161-181.

Kahneman, D. and A. Tversky (2000), *Choices, values and frames*, Cambridge: Russell Sage Foundation and Cambridge University Press.

Kahneman, D., P.P. Wakker and R. Sarin (1997), "Back to Bentham? Explorations of experienced utility", *Quarterly Journal of Economics*, Vol. 112(2), pp. 375-405.

Kapteyn, A., J.P. Smith and A. van Soest (2009), "Comparing Life Satisfaction", *RAND Labor and Population Working Paper Series*, RAND and IZA.

Keyes, C.M. (2006), "Subjective well-being in mental health and human development research worldwide: An introduction", *Social Indicators Research*, Vol. 77, pp. 1-10.

Kiecolt-Glaser, J.K., L. McGuire, T.F. Robles and R. Glaser (2002), "Psychoneuroimmunology: Psychological influences on immune function and health", *Journal of Consulting and Clinical Psychology*, Vol. 70(3), pp. 537-547.

King, G., C.J.L. Murray, J.A. Salomon and A. Tandon (2004), "Enhancing the validity and cross-cultural comparability of measurement in survey research", *American Political Science Review*, Vol. 98(1), pp. 191-207.

Kristensen, N. and E. Johansson (2008), "New evidence on cross-country differences in job satisfaction using anchoring vignettes", *Labour Economics*, Vol. 15, pp. 96-117.

Krueger, A.B. (2009), *Measuring the subjective well-being of nations: National accounts of time use and well-being*, Chicago: University of Chicago Press.

Krueger, A.B., D. Kahneman, C. Fischler, D. Schkade, N. Schwarz and A.A. Stone (2009), "Time use and subjective well-being in France and the US", *Social Indicators Research*, Vol. 93, pp. 7-18.

Lacey, H.P., D.M. Smith and P.A. Ubel (2006), "Hope I die before I get old: Mispredicting happiness across the adult lifespan", *Journal of Happiness Studies*, Vol. 7, pp. 167-182.

Larsen, R.J. and B.L. Fredrickson (1999), "Measurement issues in emotion research", in D. Kahneman, E. Diener and N. Schwarz (eds.), *Well-being. The Foundations of Hedonic Psychology*, Russel Sage Foundation, New York, pp. 40-60.

Layard, R. (2011), "Wellbeing and public policy", *CentrePiece Winter 2011/12*, CEP's Big Ideas Series, London School of Economics, Centre for Economic Performance.

Layard, R. (2005), *Happiness: Lessons from a New Science*, London: Penguin Books.

Layard, R., D. Clark, M. Knapp and G. Mayraz (2007), "Cost-benefit analysis of psychological therapy", *National Institute Economic Review*, Vol. 202, pp. 90-98.

Layard, R., G. Mayraz and S. Nickell (2008), "The marginal utility of income", *Journal of Public Economics*, Vol. 92(8-9), pp. 1846-1857.

Loewenstein, G. and P.A. Ubel (2008), "Hedonic adaptation and the role of decision and experience utility in public policy", *Journal of Public Economics*, Vol. 92, pp. 1975-1810.

Lucas, R.E. (2007a), "Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies", *Journal of Personality and Social Psychology*, Vol. 92(4), pp. 717-730.

Lucas, R.E. (2007b), "Adaptation and the set-point model of subjective well-being", *Current Directions in Psychological Science*, Vol. 16(2), pp. 75-79.

Lucas, R.E. and M. Donnellan (2011), "Estimating the Reliability of Single-Item Life Satisfaction Measures: Results from Four National Panel Studies", *Social Indicators Research*, forthcoming.

Lucas, R.E., A.E. Clark, Y. Georgellis and E. Diener (2003), "Re-examining adaptation and the set point model of happiness: Reactions to changes in marital status", *Journal of Personality and Social Psychology*, Vol. 84, pp. 527-539.

Luhmann, M., R.E. Lucas, M. Eid and E. Diener (2012), "The prospective effect of life satisfaction on life events", *Social Psychological and Personality Science*, published online before print, 20 March.

Lyubomirsky, S., L. King and E. Diener (2005), "The benefits of frequent positive affect: Does happiness lead to success?", *Psychological Bulletin*, Vol. 131(6), pp. 803-855.

McCloskey, D. (2012), "Happyism: The creepy new economics of pleasure", *The New Republic*, 8 June, available online at: *www.tnr.com/article/politics/magazine/103952/happyism-deirdre-mccloskey-economics-happiness*, last accessed 10 August 2012.

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

Metcalfe, R., N. Powdthavee and P. Dolan (2011), "Destruction and distress: Using a quasi-experiment to show the effects of the September 11 attacks on mental well-being in the United Kingdom", *The Economic Journal*, Vol. 121(550), pp. 81-103.

Moyle, P. (1995), "The role of negative affectivity in the stress process: Tests of alternative models", *Journal of Organizational Behavior*, Vol. 16(6), pp. 647-668.

Napier, J.L. and J.T. Jost (2008), "Why are Conservatives happier than Liberals?", *Psychological Science*, Vol. 19(6), pp. 565-572.

New Economics Foundation (2009), contributing authors: J. Michaelson, S. Abdallah, N. Steuer, S. Thompson and N. Marks, "National Accounts of Well-Being: Bringing real wealth onto the balance sheet", *New Economics Foundation*, January, London, available online at: *www.national-accountsofwellbeing.org/learn/download-report.html*.

Ng, Y.K. (1997), "A case for happiness, cardinalism and interpersonal comparability", *The Economic Journal*, Vol. 107, pp. 1848-1858.

Ng, Y.K. (1996), "Happiness surveys: Some comparability issues and an exploratory survey based on just perceivable increments", *Social Indicators Research*, Vol. 38, pp. 1-27.

Nussbaum, M.C. (2008), "Who is the happy warrior? Philosophy poses questions to Psychology", *The Journal of Legal Studies*, Vol. 37(2), pp. S81-S113.

OECD (2012), *Measuring regulatory performance – A practitioners guide to perception surveys*, Paris, OECD Publishing.

OECD (2011a), *How's Life? Measuring Well-Being*, Paris, OECD Publishing.

OECD (2011b), internal analysis of data shared by around 4 000 users of *Your Better Life Index*, as of September, available online at: *http://oecdbetterlifeindex.org/*.

Oishi, S. (2002), "The experiencing and remembering of well-being: A cross-cultural analysis", *Personality and Social Psychology Bulletin*, Vol. 28, pp. 1398-1406.

Oishi, S., E. Diener and R.E. Lucas (2007), "The optimum level of well-being: Can people be too happy?", *Perspectives on Psychological Science*, Vol. 2(4), pp. 346-360.

Olejnik, S. and J. Algina (2000), "Measures of effect size for comparative studies: Applications, interpretations and limitations", *Contemporary Educational Psychology*, Vol. 25, pp. 241-286.

ONS UK (2012), "Analysis of experimental subjective well-being data from the Annual Population Survey. April to September 2011", *Statistical release*, 28 February, available online at: *www.ons.gov.uk/ons/rel/wellbeing/measuring-subjective-wellbeing-in-the-uk/analysis-of-experimental-subjective-well-being-data-from-the-annual-population-survey-april-september-2011/report-april-to-september-2011.html*.

ONS UK (2011a), "Findings from the National Well-Being Debate", *UK Office for National Statistics Paper*, July 2011.

ONS 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*.

Ostir, G.V., K.S. Markides, M.K. Peek and J.S. Goodwin (2001), "The association between emotional well-being and the incidence of stroke in older adults", *Psychosomatic Medicine*, Vol. 63, pp. 210-215.

Oswald, A.J. and N. Powdthavee (2008a), "Does happiness adapt? A longitudinal study of disability with implications for economists and judges", *Journal of Public Economics*, Vol. 92, pp. 1061-1077.

Oswald, A.J. and N. Powdthavee (2008b) "Death, happiness, and the calculation of compensatory damages", *The Journal of Legal Studies*, Vol. 37(S2), pp. S217-S251.

Pavot, W. and E. Diener (1993), "Review of the Satisfaction with Life Scale", *Psychological Assessment*, Vol. 5(2), pp. 164-172.

Posner, E.A. and C.R. Sunstein (eds.) (2010), *Law and Happiness*, Chicago: University of Chicago Press.

Powdthavee, N. (2010), *The Happiness Equation*, London: Icon Books.

Powdthavee, N. (2009), "I can't smile without you: Spousal correlation in life satisfaction", *Journal of Economic Psychology*, Vol. 30, pp. 675-689.

Prentice, D.A. and D.T. Miller (1992), "When small effects are impressive", *Psychological Bulletin*, Vol. 112(1), pp. 160-164.

Pressman, S.D. and S. Cohen (2005), "Does positive affect influence health?", *Psychological Bulletin*, Vol. 131(6), pp. 925-971.

Ravallion, M. (2012), "Poor, or just feeling poor? On using subjective data in measuring poverty", *World Bank Policy Research Working Paper*, No. 5968, World Bank Development Research Group, Washington, DC.

Riis, J., G. Loewenstein, J. Baron, C. Jepson, A. Fagerlin and P.A. Ubel (2005), "Ignorance to hedonic adaptation to hemodialysis; A study using Ecological Momentary Assessment", *Journal of Experimental Psychology: General*, Vol. 134, pp. 3-9.

Riphahn, R. and O. Serfling (2005), "Item non-response on income and wealth questions", *Empirical Economics*, Vol. 30, pp. 521-538.

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., M. Lewicka and T. Niit (1989), "A Cross-Cultural Study of a Circumplex Model of Affect", *Journal of Personality and Social Psychology*, Vol. 57(5), pp. 848-856.

Ryan, R.M. and E.L. Deci (2001), "On happiness and human potentials: A review of research on hedonic and eudaimonic well-being", *Annual Review of Psychology*, Vol. 52, pp. 141-166.

Ryff, C.D. (1989), "Happiness is everything, or is it? Explorations on the meaning of psychological well-being", *Journal of Personality and Social Psychology*, Vol. 57(6), pp. 1069-1081.

Sacks, D.W., B. Stevenson and J. Wolfers (2010), "Subjective well-being, income, economic development and growth", *NBER Working Paper*, No. 16441, Cambridge, Mass.: National Bureau of Economic Research.

Schaubroeck, J., D.C. Ganster and M.L. Fox (1992), "Dispositional affect and work-related stress", *Journal of Applied Psychology*, Vol. 77, pp. 332-335.

Schkade, D.A. and D. Kahneman (1998), "Does living in California make people happy? A focusing illusion on judgments of life satisfaction", *Psychological Science*, Vol. 9(5), pp. 340-346.

Schulz, R. and S. Decker (1985), "Long-term adjustment to physical disability: The role of social support, perceived control, and self-blame", *Journal of Personality and Social Psychology*, Vol. 48(5), pp. 1162-1172.

Schwartz, P. (2012), "Happiness is not within the government's remit: The philosophical flaw in Happiness Economics", in P. Booth (ed.), *… and the Pursuit of Happiness: Wellbeing and the role of Government*, London: Institute of Economic Affairs, in association with Profile Books.

Scollon, C.N., E. Diener, S. Oishi and R. Biswas-Diener (2004), "Emotions across cultures and methods", *Journal of Cross-Cultural Psychology*, Vol. 35(3), pp. 304-326.

Sen, A. (2002), "Health: perception versus observation", *British Medical Journal*, Vol. 324, pp. 860-861.

Senik, C. (2011), "The French unhappiness puzzle: The cultural dimension of happiness", *Paris School of Economics Working Paper*, No. 2011-34, Paris-Jourdan Sciences Économiques.

Shackleton, J.R. (2012), "Wellbeing at work: Any lessons?", in P. Booth (ed.), *… and the Pursuit of Happiness: Wellbeing and the role of Government*, London: Institute of Economic Affairs, in association with Profile Books.

Smith, D., K.M. Langa, M. Kabeto and P.A. Ubel (2005), "Health, wealth and happiness: Financial resources buffer subjective well-being after the onset of disability", *Psychological Science*, Vol. 16, pp. 663-666.

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.

Statistics New Zealand (2008), available online at: *www.stats.govt.nz/browse_for_stats/environment/sustainable_development/sustainable-development.aspx*.

Steptoe, A., J. Wardle and M. Marmot (2005), "Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102(18), pp. 6508-6512.

Stevenson, B. and J. Wolfers (2008), "Economic Growth and Subjective Wellbeing: Reassessing the Easterlin Paradox", *NBER Working Paper*, No. 14282, Cambridge, Mass.: National Bureau of Economic Research.

Stiglitz, J.E., A. Sen, J.P. Fitoussi (2008), *Report by the Commission on the Measurement of Economic Performance and Social Progress*, available online at: *www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf*.

Stoll, L., J. Michaelson and C. Seaford (2012), "Well-being evidence for policy: A review", London: New Economics Foundation, available online at: *www.neweconomics.org/publications/well-being-evidence-for-policy-a-review*, last accessed 10 August 2012.

Stutzer, A. and B.S. Frey (2008) "Stress that doesn't pay: The commuting paradox", *The Scandinavian Journal of Economics*, Vol. 110(2), pp. 339-366.

Stutzer, A. and R. Lalive (2004), "The role of social work norms in job searching and subjective well-being", *Journal of the European Economic Organisation*, Vol. 2(4), pp. 696-719.

Sugden, R. (2005), "Anomalies and stated preference techniques: A framework for a discussion of coping strategies", *Environmental and Resource Economics*, Vol. 32, pp. 1-12.

Tabachnick, B.G. and L.S. Fidell (2001), *Using multivariate statistics*, fourth edition, Needham Heights, MA: Allyn and Bacon.

Thaler, R.H. and C.S. Sunstein (2008), *Nudge: Improving decisions about health, wealth and happiness*, New Haven, CT: Yale University Press.

Tsai, J.L., B. Knutson and H.H. Fung (2006), "Cultural variation in affect valuation", *Journal of Personality and Social Psychology*, Vol. 90(2), pp. 288-307.

Ubel, P.A., A. Jankovic, D. Smith, K. Langa and A. Fagerlin (2005), "What is perfect health to an 85-year-old? Evidence for scale recalibration in subjective health ratings", *Medical Care*, Vol. 43(10), pp. 1054-1057.

UMR Research (2012), "UMR Omnibus Results: Happiness of New Zealand", UMR research summary, reported by Scoop, available online at: *http://img.scoop.co.nz/media/pdfs/1202/HappinessOfNZ Update_Jan12.pdf*.

van de Vijver, F.J.R. and Y.H. Poortinga (1997), "Towards an integrated analysis of bias in cross-cultural assessment", *European Journal of Psychological Assessment*, Vol. 13(1), pp. 29-37.

Van Praag, B.M.S. and B.E. Baarsma (2005), "Using Happiness Surveys to Value Intangibles: The Case of Airport Noise", *The Economic Journal*, Vol. 115, pp. 224-246.

Veenhoven, R. (1994), "Is happiness a trait? Tests of the theory that a better society does not make people any happier", *Social Indicators Research*, Vol. 32(2), pp. 101-160.

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.

Waterman, Alan S. (2007), "On the importance of distinguishing hedonia and eudaimonia when contemplating the hedonic treadmill", *American Psychologist*, Vol. 62(6), pp. 612-613.

Wilson, T.D. and D.T. Gilbert (2005), "Affective forecasting: Knowing what to want", *Current Directions in Psychological Science*, Vol. 14(3), pp. 131-134.

White, M.P. and P. Dolan (2009), "Accounting for the richness of daily activities", *Psychological Science*, Vol. 20(8), pp. 1000-1008.

Wright, T.A. and B.M. Staw (1999), "Affect and favourable work outcomes: Two longitudinal tests of the happy-productive worker thesis", *Journal of Organizational Behavior*, Vol. 20, pp. 1-23.

Yeaton, W.H. and L. Sechrest (1981), "Meaningful measures of effect", *Journal of Consulting and Clinical Psychology*, Vol. 49(5), pp. 766-767.

Zweimüller, J. (1992), "Survey non-response and biases in wage regressions", *Economics Letters*, Vol. 39, pp. 105-109.

ANNEX A

Illustrative examples of subjective well-being measures

Example evaluative measures

i) The "Cantril Ladder", or "Cantril's Ladder of Life Scale", as adopted in the Gallup World Poll (Bjørnskov, 2010):

> *Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.*
>
> *If the top step is 10 and the bottom step is 0, on which step of the ladder do you feel you personally stand at the present time?*

ii) An overall life satisfaction question, as adopted in the World Values Survey (Bjørnskov, 2010):

> *All things considered, how satisfied are you with your life as a whole these days? Using this card on which 1 means you are "completely dissatisfied" and 10 means you are "completely satisfied" where would you put your satisfaction with life as a whole?*

iii) UK Office of National Statistics (ONS) experimental evaluative subjective well-being question, tested in the Annual Population Survey (2012) and the Opinions Survey (2011b):

> *Overall, how satisfied are you with your life nowadays?*
>
> Respondents are asked to provide an answer from 0 ("not at all") to 10 ("completely").

iv) The Andrews and Withey 1976 "Delighted-Terrible" scale (reported in Diener, 2009):

> *How do you feel about your life as a whole?*
>
> Respondents are supplied with seven response options ranging from "Delighted" to "Terrible".

Example affect measures

i) *Scale of Positive and Negative Experience* (SPANE) © Copyright by Ed Diener and Robert Biswas-Diener, January 2009. Published in E. Diener (2009), *Assessing Well-Being: The Collected Works of Ed Diener*, Dordrecht: Springer.

> *Please think about what you have been doing and experiencing during the past four weeks. Then report how much you experienced each of the following feelings, using the scale below. For each item, select a number from 1 to 5, and indicate that number on your response sheet:*

- *Positive.*
- *Negative.*

- *Good.*
- *Bad.*
- *Pleasant.*
- *Unpleasant.*
- *Happy.*
- *Sad.*
- *Afraid.*
- *Joyful.*
- *Angry.*
- *Contented.*

Response code for each item:

1. *Very rarely or never.*
2. *Rarely.*
3. *Sometimes.*
4. *Often.*
5. *Very often or always.*

ii) Huppert et al. (2009), European Social Survey well-being module:

I will now read out a list of the ways you might have felt or behaved in the past week.

Please tell me how much of the time during the past week:

a) … you felt depressed.

b) … you felt that everything you did was an effort.

c) … your sleep was restless.

d) … you were happy.

e) … you felt lonely.

f) … you enjoyed life.

g) … you felt sad.

h) … you could not get going.

i) … you had a lot of energy.

j) … you felt anxious.

k) … you felt tired.

l) … you were absorbed in what you were doing.

m) … you felt calm and peaceful.

n) … you felt bored.

o) … you felt really rested when you woke up in the morning.

Note: Items a) to h) comprise the short *Center for Epidemiologic Studies Depression Scale* (CES-D; Radloff, 1977; Steffick, 2000).

Response code for each item:

1. *None or almost none of the time.*
4. *All or almost all of the time.*

iii) UK Office of National Statistics (ONS) experimental experienced subjective well-being questions, tested in the Annual Population Survey (2012):

Overall, how happy did you feel yesterday? (positive affect)

Overall, how anxious did you feel yesterday? (negative affect)

For both items, respondents are asked to provide an answer from 0 ("not at all") to 10 ("completely).

iv) UK Office of National Statistics (ONS) extended experimental experience subjective well-being questions, tested in the August 2011 Opinions Survey (2011b).

Positive affect items:

- *Overall, how happy did you feel yesterday?*
- *Overall, how content did you feel yesterday?*
- *Overall, how calm did you feel yesterday?*
- *Overall, how relaxed did you feel yesterday?*
- *Overall, how peaceful did you feel yesterday?*
- *Overall, how much enjoyment did you feel yesterday?*
- *Overall, how joyful did you feel yesterday?*
- *Overall, how energised did you feel yesterday?*
- *Overall, how excited did you feel yesterday?*

Negative affect items:

- *Overall, how tired did you feel yesterday?*
- *Overall, how anxious did you feel yesterday?*
- *Overall, how stressed did you feel yesterday?*
- *Overall, how worried did you feel yesterday?*
- *Overall, how bored did you feel yesterday?*
- *Overall, how much pain did you feel yesterday?*
- *Overall, how angry did you feel yesterday?*
- *Overall, how lonely did you feel yesterday?*

For all items, respondents are asked to provide an answer from 0 ("not at all") to 10 ("completely").

Example eudaimonic measures

i) UK Office of National Statistics (ONS) experimental eudaimonic subjective well-being question, tested in the Annual Population Survey (2012):

Overall, to what extent do you feel the things you do in your life are worthwhile?

ii) *Psychological Well-Being Scale* (PWB) © Copyright by Ed Diener and Robert Biswas-Diener, January 2009. Published in E. Diener (2009), *Assessing Well-Being: The Collected Works of Ed Diener*, Dordrecht: Springer.

Below are 8 statements with which you may agree or disagree. Using the 1-7 response scale below, indicate your agreement with each item by indicating that response for each statement.

- *I lead a purposeful and meaningful life.*
- *My social relationships are supportive and rewarding.*

- *I am engaged and interested in my daily activities.*
- *I actively contribute to the happiness and well-being of others.*
- *I am competent and capable in the activities that are important to me.*
- *I am a good person and live a good life.*
- *I am optimistic about my future.*
- *People respect me.*

Response code for all items:

7. Strongly agree.

6. Agree.

5. Slightly agree.

4. Mixed or neither agree nor disagree.

3. Slightly disagree.

2. Disagree.

1. Strongly disagree.

iii) Huppert and So (2011), *Flourishing index*, drawn from the European Social Survey (2006/7) Round 3 supplementary well-being module.

Construct name	ESS item used as indicator	Response format
Competence	*Most days I feel a sense of accomplishment from what I do.*	5-point scale from *strongly disagree* to *strongly agree.*
Engagement	*I love learning new things.*	
Meaning	*I generally feel that what I do in my life is valuable and worthwhile.*	
Optimism	*I am always optimistic about my future.*	
Positive relationships	*There are people in my life who really care about me.*	
Resilience	*When things go wrong in my life it generally takes me a long time to get back to normal.*[1]	
Self-esteem	*In general, I feel very positive about myself.*	
Emotional stability	*(In the past week) I felt calm and peaceful.*	4-point scale from *none or almost none of the time* to *all or almost all of the time.*
Vitality	*(In the past week) I had a lot of energy.*	
Positive emotion	*Taking all things together, how happy would you say you are?*	0 to 10 scale from *extremely unhappy* to *extremely happy.*

1. Reverse-coded item.

For details of how these items were compiled into an operational definition of *flourishing*, see Huppert and So (2011).

ANNEX B

Question modules

Module A: Core measures

Objective

This module is intended to provide a minimal set of measures of subjective well-being covering both life evaluation and affect that could be included in household surveys. The core measures included here are the measures for which there is the strongest evidence for their validity and relevance, and for which international comparability is the most important. An experimental measure of an aspect of eudaimonic well-being is also included.

Description

The module contains a single question on overall life satisfaction (A1). This question is intended to capture the respondent's evaluative judgement of how their life is going while imposing the minimum level of respondent burden. It is envisaged that Question A1 will serve as the **primary measure** of subjective well-being when a single measure is required. Question A2 captures the eudaimonic concept of whether the things the respondant does in their life are worthwhile. Three questions on affect are also included (Questions A3 to A5). These should be included as a group, and are intended to provide a minimal set of questions required to characterise the affective state of the respondent on the previous day.

Box B.1. **Core questions**

The following question asks how satisfied you feel, on a scale from 0 to 10. Zero means you feel "not at all satisfied" and 10 means you feel "completely satisfied".

A1. Overall, how satisfied are you with life as a whole these days? [0-10]

The following question asks how worthwhile you feel the things you do in your life are, on a scale from 0 to 10. Zero means you feel the things you do in your life are "not at all worthwhile", and 10 means "completely worthwhile".

A2. Overall, to what extent do you feel the things you do in your life are worthwhile? [0-10]

The following questions ask about how you felt yesterday on a scale from 0 to 10. Zero means you did not experience the feeling "at all" yesterday while 10 means you experienced the feeling "all of the time" yesterday. I will now read out a list of ways you might have felt yesterday.

A3. How about happy? [0-10]

A4. How about worried? [0-10]

A5. How about depressed? [0-10]

Origin

The satisfaction with life question is based on that used in the World Values Survey, but amended to use a 0-to-10 scale. Other versions of this question have been used in the European Social Survey, the German Socio-Economic Panel, the British Household Panel Study, the Canadian General Social Survey, and more recently by the INSEE and ONS. The 0-10 response scale format and the use of "completely dissatisified" as a scale anchor have been adopted on the basis of the review of evidence in Chapter 2.

The eudaimonic question is based on the single item measure developed by the ONS in their experimental questions on subjective well-being, used in their Annual Population Survey from April 2011 to March 2012.

The affect questions used here are derived from the Gallup World Poll and the European Social Survey.

Time

The module is expected to take about 90 seconds to complete in total.

Output

Data on life satisfaction can be presented as the mean value of responses, excluding missing values. Standard measures of distribution used should be the standard deviation of responses and the inter-quartile range of responses. The mean value of responses, and the standard error of this estimate, could be used to describe differences in life satisfaction among sub-groups of the population. The percentage of the population reporting a life satisfaction below a "low life satisfaction" threshold could also be usefully reported.

The experimental eudaimonic measure (A2) should be reported in a similar way to the primary measure of life satisfaction (A1).

Information from the affect questions can be presented either as the results of answers to single questions or as a composite index. The answers to individual questions provide information on particular emotional states. The composite measures can be used to summarise negative affect and the respondent's affect balance.

Information on responses to individual questions (A3 to A5) can be presented as the mean value of responses, excluding missing values. Standard measures of distribution used should be the standard deviation of responses and the inter-quartile range of responses. The mean value of responses, and the standard error of this estimate, could be used to describe differences in affect among sub-groups of the population. The percentage of the population reporting a "low" level of a particular mood (below threshold) could also be usefully reported.

Only a single positive affect question is included here A3 (happy), so there is no need to construct a positive affect index.

A composite measure of negative affect is calculated as the average score for Question A4 (worried) and Question A5 (depressed). This will give a value in the 0-to-10 range.

Affect balance can be calculated as positive affect less negative affect for each individual respondent and averaged across respondents. This will give a value ranging from -10 to 10. Affect balance can be reported as the mean score (-10 to 10) but also as the proportion of the population with net negative affect (an affect balance less than 0), sometimes described as a U-index (Kahneman and Krueger, 2006).

Guidelines for interviewers

The primary question (A1) deliberately focuses on how people are feeling "these days" rather than specifying a longer or shorter time period. The intent is not to obtain the current emotional state of the respondent, but to obtain a cognitive evaluation on their level of life satisfaction.

The experimental eudaimonic question (A2) concerns the extent to when people feel their activities in general ("the thing you do in your life") are worthwhile. No specific time frame is supplied: respondents are invited to make an overall assessment.

Questions A3 through to A5 focus on the respondent's moods and feelings on the previous day. The time-frame is explicitly short because the primary focus is the feelings that people actually experienced, not an overall assessment of how things are going these days. If a respondent indicates that the previous day was unusual in some respect (something particularly bad or good happened, or they were feeling unwell), they should still report how they felt that day. We are interested in the feelings people have actually experienced, not how people feel on a "typical" day. Because a large number of people are being interviewed over a relatively long period of time, unusual events will not overly bias the aggregated statistics that are produced. More importantly, the reference to a specific day permits the data to be used to unravel day-of-week effects and responses to external events for which the dates are known.

Module B. Life evaluation

Objective

This module contains questions on the respondent's cognitive judgements on how they evaluate their own lives. It is not intended to be used in its entirety, but provides a range of possible life evaluation measures that are complementary to the primary measure described in Module A. These measures could be used in circumstances where a more in-depth understanding of subjective well-being is required or to help understand methodological issues in measuring subjective well-being. In all cases, these questions should come after those in Module A.

Description

The life evaluation module contains a range of different questions filling different purposes. Questions B1 (the self-anchoring striving scale or "Cantril Ladder") and B2 (overall happiness) are alternative ways of measuring the same underlying construct as the primary measure of life evaluation included in the Core module (A1). These may be used to complement the existing primary measure where additional measures would add value.

Questions B3 and B4 capture information on respondent's perceptions of their life satisfaction in the past and anticipated in the future. This provides information about how optimistic or pessimistic the respondent is, but also adds information on the respondent's overall life evaluation, as people's expectations of the future are part of how they evaluate their life.

Questions B5 to B9 together define the *Satisfaction With Life Scale* (SWLS). The SWLS is one of the best-tested and most reliable multi-item scales of life evaluation, has a higher reliability than single item measures, and is more robust to inter-personal differences in scale interpretation than a single-item measure. It should be noted, however, that this is not a balanced scale, and there is a slightly increased risk of acquiescence/socially desirable responding due to the use of an agree/disagree scale format.

Box B.2. **Life evaluation questions**

Please imagine a ladder with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.

B1. On which step of the ladder would you say you personally feel you stand at this time? [0-10]

The following question asks how happy you feel, on a scale from 0 to 10. Zero means you feel "not at all happy" and 10 means "completely happy".

B2. Taking all things together, how happy would you say you are? [0-10]

The following questions ask how satisfied you feel, on a scale from 0 to 10. Zero means you feel "not at all satisfied" and 10 means "completely satisfied".

B3. Overall, how satisfied with your life were you 5 years ago? [0-10]

B4. As your best guess, overall how satisfied with your life do you expect to feel in 5 years time? [0-10]

Below are five statements with which you may agree or disagree. Using the 1-7 scale below, indicate your agreement with each item. Please be open and honest in your responding. The 7 point scale is as follows:

1. *Strongly disagree.*
2. *Disagree.*
3. *Slightly agree.*
4. *Neither agree nor disagree.*
5. *Slightly agree.*
6. *Agree.*
7. *Strongly agree.*

B5. In most ways my life is close to my ideal[1-7]

B6. The conditions of my life are excellent [1-7]

B7. I am satisfied with my life [1-7]

B8. So far I have gotten the important things I want in life [1-7]

B9. If I could live my life over, I would change almost nothing [1-7]

Origin

The Cantril Ladder (B1) was developed by Hadley Cantril in 1961 and has been widely used subsequently. The overall happiness question used here (B2) is based on that used in the European Social Survey 2006/7. Questions B3 and B4 are based on those used by the ONS in the June 2011 Opinion survey, but using the scale anchor "completely dissatisfied", based on the recommendations of Chapter 2. The SWLS was developed by Ed Diener and William Pavot in 1985, and is used without modification.

Time

Individual Questions B1 to B4 are expected to take around 20 seconds or less each to complete. The group of questions from B5 to B9 are expected to take around 2 minutes to complete in total.

Output

Data on Questions B1 through to B4 can be disseminated as the mean value of responses, excluding missing values. The main measures of distribution for these measures should be the standard deviation of responses and the inter-quartile range of responses. The mean values of responses, and the standard errors of the means, can be used to analyse differences among sub-groups of the population. The percentage of the population reporting a life satisfaction below a "low life satisfaction" threshold could also be usefully reported.

The Satisfaction With Life Scale is calculated as the sum of responses to each of the questions from B5 through to B9. This gives a score of 5 to 35. The mean score should be output, along with the standard deviation as a measure of distribution. A score of 20 represents the neutral point on the scale. Descriptive labels have been attached to mean scores as follows (Pavot and Diener, 1992):

5 -9 Extremely dissatisfied.

10-14 Dissatisfied.

15-19 Slightly dissatisfied.

20 Neither satisfied nor dissatisfied.

21-25 Slightly satisfied.

26-30 Satisfied.

31-35 Extremely satisfied.

Guidelines for interviewers

These questions deliberately focus on how people are feeling about life as a whole rather than specifying a longer or shorter time period and ask the respondents for a reflective judgement rather than a statement of their current mood. Questions B3 and B4 ask respondents to reflect on their judgement of life as a whole five years in the past and how they see it five years in the future.The intent is not to obtain the current emotional state of the respondent, but for them to make a reflective judgement on their level of satisfaction.

Module C. Affect

Objective

This module is intended to collect information on recent positive and negative emotional states. The questions on positive and negative affect capture distinct aspects of subjective well-being that are not reflected in more evaluative measures.

Description

This module includes ten questions on affect yesterday. Questions C1 through to C4 repeat Questions A2 through to A5. They are repeated here in the event that an affect module is included in a survey that does not include affect questions among its core measures. If the core measures are included in full, Questions C1 to C4 can be dropped.

Questions C1, C2, C5 and C10 capture aspects of positive affect. Questions C3, C4, C6, C7, C8 and C9 capture aspects of negative affect. There are more negative than positive questions, reflecting the fact that negative affect is intrinsically more multi-dimensional than positive affect. Questions C2, C4, C6 and C9 can be considered as capturing states of low arousal, while the remaining questions can be considered measures of states of high arousal.

Box B.3. **Affect questions**

The following questions ask about how you felt yesterday on a scale from 0 to 10. Zero means you did not experience the emotion "at all" yesterday while 10 means you experienced the emotion "all of the time" yesterday. I will now read out a list of ways you might have felt yesterday.

C1. How about enjoyment?	[0-10]
C2. How about calm?	[0-10]
C3. How about worried?	[0-10]
C4. How about sadness?	[0-10]
C5. How about happy?	[0-10]
C6. How about depressed?	[0-10]
C7. How about anger?	[0-10]
C8. How about stress?	[0-10]
C9. How about tired?	[0-10]
C10. Did you smile or laugh a lot yesterday?	[0-10]

Origin

The affect questions used here are derived from the Gallup World Poll and the European Social Survey. Minor wording changes have been made on the basis of Chapter 2.

Time

This module is expected to take about 3 minutes to complete.

Output

Information from the affect questions in this section can be presented either as the results of answers to single questions or as a composite index. The answers to individual questions provide information on particular emotional states. The composite measures capture aspects of positive affect, negative affect or the respondent's affect balance.

Information on responses to individual questions can be presented as the proportion of respondents indicating that they experienced the relevant feeling a lot yesterday.

A composite measure of positive affect can be calculated as the mean score for Questions C1, C2, C5 and C10. This will give a value in the 0-to-10 range.

A composite measure of negative affect can be calculated as the mean score for Questions C3, C4, C6, C7, C8 and C9. This will give a value in the 0-to-10 range.

A composite measure of affect balance can be calculated as positive affect minus negative affect for each respondent averaged across all respondents. This will give a value ranging from -10 to 10. Affect balance can be reported as the mean score (-10 to 10) but also as the proportion of the population with net negative affect (an affect balance less than 0), sometimes described as a U-index (Kahneman and Krueger, 2006).

In cleaning and preparing affect data, it is important to screen for response sets. These can be most easily detected when the respondent responds 10 or 0 consistently to all 10 questions, which may indicate a lack of understanding on the part of the respondent, or an unwillingness to respond meaningfully. In either case, the lack of variation will distort subsequent analysis. Hence, such responses (where the respondent gives the same score

for all ten questions) should be coded as missing data. Whilst this procedure cannot correct for the more subtle influences of response sets/social desirability biases, it can remove the most obvious data distortions.

Guidelines for interviewers

The aim of this set of questions is to capture information on the respondent's moods on the previous day. The time-frame is explicitly short because we are interested in the feelings that people actually experienced, not an overall assessment of how things are going these days. If a respondent indicates that the previous day was unusual in some respect (something particularly bad or good happened, or they were feeling unwell), they should still report how they felt that day. We are interested in the feelings people have actually experienced, not how people feel on a "typical" day. Because we are interviewing a large number of people, we can expect that unusual events will not overly bias the aggregated statistics that are produced.

Module D. Eudaimonic well-being

Objective

This module contains questions on different aspects of people's psychological functioning. It aims to measure a range of different concepts that are sometimes described as the "eudaimonic" (or "Aristotelian") notions of well-being.

Description

The questions in the eudaimonic well-being block are relatively diverse and cover a range of different mental attributes and functionings that are thought to constitute mental "flourishing". These questions are organised into two different groups. The first set of questions (D1 to D6) is on the degree to which respondents agree or disagree with various statements about themselves, while the second set of questions (D7 to D9) is more experiential in nature. This distinction is grounded in how the questions are asked and does not represent an underlying conceptual distinction relating to different elements of eudaimonia. While individual questions may be used in different contexts if there is an identified policy need, the module does not have distinct question sub-groups intended for different purposes.

If the responses to Questions D1 to D9 are to be summed, it should be noted that collectively the items do not offer a balanced scale (with only two negatively-keyed items), and there may be a slightly increased risk of acquiescence/socially desirable responding due to the use of an agree/disagree scale format. Further development of this measure in the future is desirable.

Origin

The eudaimonia module proposed here is based on elements of the European Social Survey well-being module and the *Flourishing Scale* proposed by Diener et al. (2010). The scale anchor "did not experience this feeling" for items D1-D6 has been changed to "disagree completely", to ensure consistency between the scale anchors. Items D4 and D8 have also been changed from double-barrelled questions "valuable and worthwhile/calm and peaceful" to single-item measures, so that it is clear which word respondents are reacting to.

Box B.4. **Eudaimonic questions**

I now want to ask you some questions about how you feel about yourself and your life.

Please use a scale from 0 to 10 to indicate how you felt. Zero means you "disagree completely" and 10 means "agree completely".

D1. In general, I feel very positive about myself [0-10]

D2. I'm always optimistic about my future [0-10]

D3. I am free to decide for myself how to live my life [0-10]

D4. I generally feel that what I do in my life is worthwhile [0-10]

D5. Most days I get a sense of accomplishment from what I do [0-10]

D6. When things go wrong in my life it generally takes me a long time to get back to normal [0-10]

I am now going to read out a list of ways you might have felt during the past week. On a scale from 0 to 10, where zero means you felt that way "not at all" during the past week and 10 means you felt that way "all the time" yesterday, can you please tell me how much of the time yesterday…

D7. … you had a lot of energy? [0-10]

D8. … you felt calm? [0-10]

D9. … you felt lonely? [0-10]

Time

This module can be expected to take 4 minutes to complete.

Output

The nine questions contained in the eudaimonic well-being block are intended to be used independently in order to investigate different aspects of eudaimonic well-being. There is currently no generally accepted multi-item measure of eudaimonic well-being, although several have been proposed in the literature (Huppert and So, 2008).

Information on individual questions can be presented as the mean value of the responses, omitting missing values. Standard measures of distribution used should be the standard deviation of responses and the inter-quartile range of responses. The mean value of responses, and the standard error of this estimate, could be used to describe differences in responses to the various questions among sub-groups of the population. The percentage of the population reporting below a "poor outcomes" threshold could also be usefully reported.

In cleaning and preparing eudaimonia data, it is important to screen for response sets. These can be most easily detected when the respondent scores at the top or bottom of the scale for all nine measures, which may indicate a lack of understanding on the part of the respondent, or an unwillingness to respond meaningfully. In either case, the lack of variation will distort subsequent analysis. Hence, such responses (where the respondent gives an identical score for all questions) should be coded as missing data. Whilst this procedure cannot correct for the more subtle influences of response sets/social desirability biases, it can remove the most obvious data distortions.

Guidelines for interviewers

Questions D1 to D6 assess the degree to which respondents agree or disagree with various statements about themselves. These questions are intended to capture how people see themselves rather than emotions or feelings they have experienced. As a result, the questions are about how people are now, and do not refer to a specific time period.

The second set of questions (D7 to D9) is more experiential in nature. These questions ask about whether people actually experienced the indicated feelings during the previous week.

There is some evidence from cognitive testing that a small proportion of respondents – particularly if they are unemployed or disabled – may experience some distress in answering some questions in this module. Support and training should be provided to interviewers to enable them to respond to these circumstances appropriately.

Module E. Domain evaluation

Objective

The domain evaluation module aims to collect people's evaluative judgements on how well different aspects of their life are going using a similar question format and structure to the question on overall life satisfaction used in the Core measures (Module A). The measures presented here are intended to capture people's *satisfaction* with respect to particular domains, and are not intended as subjective measures of that domain itself (for example, in the case of the health domain, the question asks "how satisfied are you with your health status" rather than "how would you describe your health status").

Description

This module contains ten example questions on satisfaction with different aspects of life. Each life domain can be potentially analysed in its own right. In addition, the measures can be summed to calculate a composite index, along the lines of the *Personal Wellbeing Index* (PWI; International Wellbeing Group, 2006), a measure which includes items E1, E2, E3, E4, E5, E6, E7. Questions E8, E9 and E10 are not included in the PWI, but may be of relevance to policy users, or in monitoring specific aspects of well-being in their own right.

Origin

The questions used in this module are derived from the *Personal Wellbeing Index* or PWI (International Wellbeing Group, 2006) and the domain satisfaction questions used by the ONS in their June 2011 opinion survey. Use of the scale anchor "not at all satisfied" is preferred to other alternatives in the literature.

Time

This module takes 3 minutes to complete.

Output

Information on individual questions can be presented as the mean value of each response, omitting missing values. Standard measures of distribution used should be the standard deviation of responses and the inter-quartile range of responses. The mean value of responses, and the standard error of this estimate, could be used to describe differences among sub-groups of the population. The percentage of the population reporting scores below a "low satisfaction" threshold could also be usefully reported.

Box B.5. **Domain evaluation questions**

The following questions ask how satisfied you feel about specific aspects of your life, on a scale from 0 to 10. Zero means you feel "not at all satisfied" and 10 means "completely satisfied".

E1. How satisfied are you with your standard of living?	[0-10]
E2. How satisfied are your with your health?	[0-10]
E3. How satisfied are you with what you are achieving in life?	[0-10]
E4. How satisfied are you with your personal relationships?	[0-10]
E5. How satisfied are you with how safe you feel?	[0-10]
E6. How satisfied are you with feeling part of your community?	[0-10]
E7. How satisfied are you with your future security?	[0-10]
E8. How satisfied are you with the amount of time you have to do the things that you like doing?	[0-10]
E9. How satisfied are you with the quality of your local environment?	[0-10]
For respondents who are employed only:	
E10. How satisfied are you with your job?	[0-10]

The *Personal Wellbeing Index* can be calculated as the mean score of Questions E1 to E7. Missing values should be omitted. Once calculated, the *Personal Wellbeing Index* has been interpreted as a multi-item measure of overall life evaluation, albeit quite different in nature and structure to those described in Modules A and B.

In cleaning and preparing domain satisfaction data, it is important to screen for response sets. These can be most easily detected when the respondent scores at the top or bottom of the scale for all aspects (9 or 10, depending on how many were asked). This may indicate either a lack of understanding on the part of the respondent or an unwillingness to respond meaningfully. In either case, the lack of variation will distort subsequent analysis. Hence, such responses (where the respondent gives a consistent maximum or minimum score for all domain satisfaction questions) should be coded as missing data. Whilst this procedure cannot correct for the more subtle influences of response sets/social desirability biases, it can remove the most obvious data distortions.

Guidelines for interviewers

In this series of questions, respondents are being asked to make a series of evaluative judgements about different aspects of their life. Respondents should try not to let judgements about one aspect of their life affect evaluations of other aspects.

Module F. Experienced well-being

Objective

This question module focuses on questions that could be included in a time-use survey. It outlines approaches to collecting information on the positive and negative emotional states that people experienced while undertaking specific activities.

Description

The experienced well-being module has two components. The first component, comprising Questions F1 to F7, is an implementation of the Day Reconstruction Method (DRM) adapted for large-scale time-use surveys. These questions should be used together in the manner described below, and in conjunction with a time-use diary. The questions are repeated for three randomly selected time-use diary episodes.

Box B.6. **Day reconstruction method questions**

I now want to ask you some questions about how you felt yesterday.

The computer has selected three time intervals from your diary that I will ask you about.

[For each episode:]

Between ***[start time of episode]*** *and* ***[end time of episode]*** *yesterday, you said you were doing* ***[activity]****. The next set of questions asks you how you felt during this particular time.*

The following questions ask how you feel about yourself and your life, on a scale from 0 to 10. Zero means you did not experience the emotion "at all" at that time while 10 means you experienced the emotion "a lot"" at that time.

F1. Overall, how happy did you feel during this time?	[0-10]
F2. Overall, how calm did you feel during this time?	[0-10]
F3. Overall, how angry did you feel during this time?	[0-10]
F4. Overall, how sad did you feel during this time?	[0-10]
F5. Overall, how much pain did you feel during this time?	[0-10]
F6. Overall, how tired did you feel during this time?	[0-10]
F7. Were you interacting with anyone during this time, including over the phone?	[yes/no]

If yes, with whom were you interacting? *[include people on the telephone/online chat, etc.]*

Note: [**Activity**] refers to the respondent's primary activity for the episode being discussed.

The second part of the module consists of a single question (F8), which is also intended to be used as part of a time-use diary. Question F8 should generally not be used in conjunction with the DRM, as it is a substitute, and should be completed by the respondent for all time-use diary activities.

Origin

The version of the DRM used here is taken from the American Time Use Survey 2011. Question F8 was taken from the *Enquête Emploi du temps 2011*. Questions remain unaltered.

Time

The DRM is expected to take 5 to 10 minutes to complete for three activities. Question F8 is expected to add an extra 5 minutes to the time it takes respondents to complete their time-use diary, but has no effect on interview time.

Box B.7. **Experienced well-being question**

Question F8 below should be included in the time-use diary filled out by respondents. See below for an example.

F8. Was this moment pleasant or unpleasant? [from -3: very unpleasant to +3: very pleasant]

	Qu'avez-vous fait durant les 3 heures qui ont précédé la visite de l'enquêteur ?	Faisiez-vous autre chose en même temps ?	Était-ce un moment agréable ou désagréable ? (de – 3 : très désagréable à + 3 : très agréable)
..... h 00			– 3 – 2 – 1 0 + 1 + 2 + 3
10			– 3 – 2 – 1 0 + 1 + 2 + 3
20			– 3 – 2 – 1 0 + 1 + 2 + 3
30			– 3 – 2 – 1 0 + 1 + 2 + 3
40			– 3 – 2 – 1 0 + 1 + 2 + 3
50			– 3 – 2 – 1 0 + 1 + 2 + 3
..... h 00			– 3 – 2 – 1 0 + 1 + 2 + 3

Comments

The DRM (Questions F1 to F7) should be administered in an interview following the completion of a time-use diary. Because recall is important, it is desirable that the interview take place as soon as possible after the diary has been completed – preferably the day after the day covered by the diary. Question F7 relates to who the respondent was with at the time of the activity, and is conceptually distinct from the affect questions (F1 to F6). If the time-use survey already collects "who with" information, Question F7 can be omitted.

When implementing the question module, three episodes are selected from the time-use diary, omitting episodes when the respondent was sleeping or otherwise unconscious. The procedure to select the episodes should ensure that, over the sample as a whole, there are an adequate number of responses for each major time-use activity. The classification of activities can be drawn from the standard time-use classifications underpinning the survey. The questions are administered to the respondent with respect to each of the three episodes.

Question F8 is included in the time-use diary that the respondent completes rather than being administered in a follow-on interview.

Output

Information from the DRM questions described here (F1 to F7) can be presented both as the results of answers to single questions or as a composite measure of affect balance by activity classification. The answers to individual questions provide information on particular emotional states. The composite measures capture aspects of the respondent's affect balance – positive mood, negative mood, and which of the two is the stronger. In all cases, the answers should be presented with respect to a particular activity.

Information on responses to individual questions can be presented as the mean value of responses, excluding missing values for a particular activity. This will give a value in the 0-to-10 range.

A composite measure of positive affect can be calculated as the average score for Question F1 (happy) and Question F2 (calm), excluding missing values. This will give a value in the 0-to-10 range.

A composite measure of negative affect can be calculated as the average score for Questions F3 (angry), F4 (sad), F5 (pain) and F6 (worry), excluding missing values. This will give a value in the 0-to-10 range.

A composite measure of affect balance can be calculated as the difference of positive affect less negative affect for each respondent divided by 6 and averaged over all respondents. This will give a value ranging from -10 to 10. Affect balance can be reported as the mean score (-10 to 10), but can also usefully be presented as the proportion of the population with net negative affect (an affect balance less than 0), sometimes described as a U-index (Kahneman and Krueger, 2006).

In cleaning and preparing affect data, it is important to screen for response sets. These are evident when the respondent scores at the top or bottom of the scale for all six affect measures. This may indicate a response set due to either a lack of understanding on the part of the respondent or an unwillingness to respond meaningfully. In either case, the lack of variation will distort subsequent analysis. Hence, such responses (where the respondent gives the same score for all six affect questions) should be coded as missing data.

Information from the "pleasant/unpleasant" approach (Question F8) is conceptually similar to affect balance calculated from DRM data, as discussed in the previous paragraphs. Responses to Question F7 can be presented as the mean score for different activity types or the mean score for different demographic groups (e.g. sex, age groups, labour force status).

Guidelines for interviewers

These questions relate to how the respondent felt during a specific episode identified from a time-use diary. It is important that the respondent answers with respect to how they felt during the period of time covered by that episode rather than providing information on how they felt during the day as a whole or what the dominant emotion was during the day.

For Question F5, pain includes both physical and mental pain.

For Question F7, interacting means communicating or responding to someone in some way. This could include both active participation in a conversation, listening to a conference call, or playing a game like tennis or chess.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

OECD PUBLISHING, 2, rue André-Pascal, 75775 PARIS CEDEX 16
(30 2013 03 1 P) ISBN 978-92-64-19164-8 – No. 60521 2013

OECD Guidelines on Measuring Subjective Well-being

Being able to measure people's quality of life is fundamental when assessing the progress of societies. There is now widespread acknowledgement that measuring subjective well-being is an essential part of measuring quality of life alongside other social and economic dimensions. As a first step to improving the measures of quality of life, the OECD has produced *Guidelines* which provide advice on the collection and use of measures of subjective well-being. These *Guidelines* have been produced as part of the OECD *Better Life Initiative*, a pioneering project launched in 2011, with the objective to measure society's progress across eleven domains of well-being, ranging from jobs, health and housing, through to civic engagement and the environment.

These *Guidelines* represent the first attempt to provide international recommendations on collecting, publishing, and analysing subjective well-being data. They provide guidance on collecting information on people's evaluations and experiences of life, as well as on collecting "eudaimonic" measures of psychological well-being. The *Guidelines* also outline why measures of subjective well-being are relevant for monitoring and policy making, and why national statistical agencies have a critical role to play in enhancing the usefulness of existing measures. They identify the best approaches for measuring, in a reliable and consistent way, the various dimensions of subjective well-being, and provide guidance for reporting on such measures. The *Guidelines* also include a number of prototype survey modules on subjective well-being that national and international agencies can use in their surveys.

OECD*publishing*
www.oecd.org/publishing

With the financial assistance
of the European Union

ISBN 978-92-64-19164-8
30 2013 03 1 P

9 789264 191648