*Oxford Economic Papers*, 68(4), 2016, 861–870
doi: 10.1093/oep/gpw022
Advance Access Publication Date: 12 June 2016
OXFORD

# Understanding context effects for a measure of life evaluation: how responses matter

By Angus Deaton[a] and Arthur A. Stone[b]

[a]Center for Health and Wellbeing, Woodrow Wilson School, Princeton University, Princeton, NJ, USA; e-mail: deaton@princeton.edu
[b]Department of Psychology and USC Dornsife Center for Self-Report Science, University of Southern California

## Abstract

We study context effects on responses to wellbeing questions. We find that those who were randomized into being asked a series of political questions subsequently report lower life evaluation; those who were previously asked about their evaluation of the direction of the United States lowered their own life evaluation, but only if they disapproved of the way the country was going. Subgroups of the population are affected in different ways; the age profile of wellbeing is tipped in favor of the elderly, and African American's life evaluations are increased when they are asked about President Obama's performance. The context effects are large, not easily removed, and change wellbeing rankings across groups.

**JEL classification**: I31

## 1. Introduction and background

There is currently great interest in and increasing use of self-reported measures of subjective well-being (SWB), sometimes loosely referred to as measures of 'happiness'. In fact, there are at least three types of SWB measures that tap different aspects of the construct; they are eudaimonic (meaning and purpose), evaluative (satisfaction with life), and hedonic or experiential (everyday joys and pains) (Kahneman *et al.*, 1999). Evaluative SWB is the concept most commonly measured by national statistical offices (e.g. the Office of National Statistics in the UK) and in international surveys (e.g. the Heath and Retirement Study in the USA), and in this study we examine context effects on the Cantril ladder measure of evaluative SWB.

Context effects are defined as effects of preceding items or experiences on responses to subsequently presented items, and they have been known to survey methodologists and behavioural scientists for decades (Sudman *et al.*, 1996; OECD, 2013). Two pathways may explain how context effects occur, including a shift in the effect caused by the context items (negative mood causing subsequent questions to be rated more negatively) and a shift in attention to particular experiences caused by the context items (Oishi *et al*., 2003; Schwarz and Strack, 1999; Oishi *et al*. 2003). Regarding the second mechanism, respondents might unconsciously pick up cues that indicate what the interviewer is looking for, they might focus on the topic raised by earlier questions when thinking about their lives, their answers may be shaded by their current mood, or they may take little mental effort to answer the question (see also Bertrand and Mullainathan, 2001).

Items that are difficult to answer, because they require reaching back into vague remembrances or because they tap topics that are difficult to evaluate, are especially prone to context effects (Schwartz and Strack, 1999). When responding to such items, people may use whatever information comes to mind, including information provided by the immediate context. Evaluative SWB questions likely qualify as 'difficult' items given the cognitive effort required to evaluate and summarize one's life; therefore, fully understanding the impact of context on evaluative SWB is especially important.

The goal of this paper is to advance understanding of context effects by extending analyses of a unique, large-scale experiment conducted by the Gallup Organization that tested the impact of political questions (the context) preceding a question on SWB (the Cantril ladder). We use data from the Gallup–Healthways Well-being Index poll, a telephone poll of 1,000 Americans each day, which contains an evaluative well-being question. The poll was begun in January 2008 and contains the Cantril ladder question (Cantril, 1965). This asks people where they stand on an 11-point scale from 0, 'the worst possible life for you', to 10, 'the best possible life for you', which is interpreted as overall evaluation of their well-being. Gallup uses the same daily poll for its political questions, which asks about voting intentions, evaluations of the president, and assessments of the state of the country. The Gallup interview, as originally designed, began with these political questions, which change from time to time, immediately followed by the Cantril ladder. The probable existence of a context effect was revealed when it was noticed that on the few days when there were no political questions—for example, on President Obama's inauguration day or on other days where there was no obvious salient event—the mean response to the ladder was unusually high, and higher than any effect that could be reasonably attributed to that event or to events on other days when there were no political questions.

To investigate further, Gallup ran a randomized controlled trial within the ongoing survey, dividing their respondents into 500 who received the political questions and 500 who did not. The results confirmed that asking the political questions caused the average ladder score to fall by 0.67 rungs, a substantial effect that is, for comparison, as large as the effects of the financial crisis on well-being, or as large as the effect of making everyone unemployed (for details see Deaton, 2012). Despite all that is known about context effects, it is surprising that the Gallup Organization, a high-quality and experienced polling organization, did not immediately realize the pernicious impact of the political questions on the ladder. It is also to their credit that they explicitly investigated the possible effects and later altered the position of the SWB question to avoid context effects of prior questions. Indeed, in their most recent polls, they have asked the political and well-being questions in different surveys with different respondents.

In this paper, we examine the context effect induced by the political questions in more detail and go beyond earlier work by Deaton (2012) that established its existence and size, but not the mechanism. Without any direct evidence, Deaton conjectured that the context effects worked because simply being asked the context questions negatively affected people's mood, which then influenced responses to the subsequent question on well-being. We take up this question more seriously here and explore a different possibility, that it is the actual *answers* that people give to the political questions that is predictive of later effects on the Cantril ladder, not simply the fact of being asked the *questions*. If it is the answers that matter, the size of the context effect should vary across different answers to the previous questions, raising the possibility that the mean responses of different groups of respondents are affected differently, something that we investigate. To our knowledge, this group effect has not previously been reported in either economics or psychology and it could shed light on the meaning of context effects. Beyond that, such effects could have serious consequences for interpreting the between-group differences, as we show below.

Our method here is a subgroup analysis within Gallup's randomized controlled trial. We can only analyse issues that arose naturally in Gallup's work, so we cannot do any manipulation of our own, which limits the range of issues we can address. An offsetting advantage is that we have a much larger number of observations than would be possible in a laboratory study, at least at reasonable cost; that we are working with a sample that is representative of the US adult population; and that we are looking at responses that actually took place in real polling not among those brought into a laboratory.

Our results show that, although politics and politicians in the USA were widely unpopular at the time of the survey, it is not simply the asking of these questions that resulted in lower well-being. Instead, it is only those who answer negatively who lower their own self-reported well-being. As argued by Schwarz and Strack, this suggests that respondents reach back to easily available information on the topic when they answer the well-being question. They thus give more weight to the state of the country, the question immediately preceding the ladder, when assessing their own well-being than would have been the case had their opinion of the state of the country not been recently in mind. There is no such effect for those who are satisfied with the direction of the country. The context effects, although large, do not much affect standard patterns of life evaluation by sex, age, or income group. There is one exception that is readily explained and perhaps even predictable, but again demonstrates the power of previous questions. African Americans rate their well-being more highly than do either Whites or Hispanics, but only after they have been asked about the performance of President Obama. Otherwise, they are the group with the lowest life evaluation.

## 2. Data and experimental evidence

The data we analyse come from the period during which Gallup was researching the size of the context effect and had randomly split the daily sample of 1,000 respondents, with 500 people being asked the political questions followed by the ladder question and 500 receiving no political questions, starting instead with the ladder question. Table 1 shows the sequence of questions for the treatment group; the ladder follows the political questions and, for the control group, the ladder is the first question.[1]

Table 2 documents the average treatment effect (ATE) of being asked the prior questions on a range of outcomes, as well as the ladder, which is our main focus: we consider the

[1] We note that Gallup altered the political questions and inserted a 'buffer' question to reduce context effects on the ladder (see Deaton, 2012), but we focus exclusively in the single period described in Table 1 to support our argument.

**Table 1.** Question order for three experimental periods

| Question order | 21 January 2009 to 5 April 2009 (75 days, 76,167 observations) |
|---|---|
| ↓ | Do you approve or disapprove of the way that Barack Obama is handling his job as president? |
| | In general, are you satisfied or dissatisfied with the way things are going in the USA? |
| | Cantril ladder |

**Table 2.** Effect of prior questions on subsequent questions

| Variables | Mean over treatments and controls | Average treatment effects | |
|---|---|---|---|
| | | ATE | *t* |
| Ladder | 6.45 | **−0.67** | (24.0) |
| Health is excellent | 0.207 | 0.011 | (3.3) |
| Health is very good | 0.287 | −0.009 | (2.3) |
| Health is good | 0.296 | −0.009 | (2.2) |
| Health is fair | 0.150 | −0.001 | (0.2) |
| Health is poor | 0..058 | **0.008** | (3.5) |
| Disability? | 0.220 | −0.002 | (0.6) |
| Smoker? | 0.215 | 0.002 | (0.7) |
| Personal doctor? | 0.802 | 0.001 | (0.4) |
| SOL is OK | 0.725 | **−0.033** | (7.7) |
| SOL going up | 0.364 | **−0.029** | (6.0) |
| SOL same | 0.211 | **−0.015** | (4.3) |
| SOL falling | 0.425 | **0.044** | (9.9) |
| Married? | 0.519 | −0.010 | (2.1) |
| Hispanic? | 0.111 | −0.001 | (0.4) |
| High income? | 0.351 | −0.008 | (2.1) |
| Income refused or don’t know | 0.206 | 0.005 | (1.5) |

*Notes:* The ladder is on a scale from 0 to 10, other variables are 0 or 1. High income is an indicator that monthly income was declared to be at least $4,000. The self-assessed health, disability, smoking, and whether or not you have a personal doctor closely follow the ladder questions in the questionnaire. The standard-of-living (SOL) questions follow the health questions. SOL OK is 1 if respondent says SOL is satisfactory and SOL going up, same, or falling are three answers to a question about whether the SOL is getting better, staying the same, or getting worse; the coefficients on those three add to zero. Absolute *t*-values are shown in brackets. The bold ATEs are those with *t*-values greater than or equal to 3.5, which is close to the square root of the logarithm of the sample size, the BIC or Schwarz (1978) large-sample Bayesian test that adjusts for sample size.

effects on reports of health-related, demographic, and financial perception variables. The first column shows the means (over both treatments and controls, i.e. form 1 and form 2 respondents). On average, people rate their lives highly, at 6.45 on the 0 to 10 scale, with a standard deviation of 2.12. There is a very large, −0.67, reduction in the reported ladder among those who were randomized into receiving the political questions. The second and third columns list the estimated ATEs and the *t*-values for the hypotheses that each ATE is zero. Given the large sample sizes here, a better trade-off between type I and type II errors is given by the Schwarz (1978) or Bayesian information criterion (BIC) test, which is that the *F*-statistic exceeds the logarithm of the sample size; such cases are shown in bold in the table.

With one exception, there are no other ATEs that are close to being as large as the ATE on the ladder, whether judged by the fraction of the mean response or by statistical significance. The exception is the standard-of-living question, where there is a reduction of 3.3 percentage points from the mean, and there is a 4.4 percentage point increase in those who say they expect their standard of living to fall, which is even larger than the effect on the ladder relative to the mean.

Once again, it is plausible that the question about whether the respondent approves of the 'way things are going in the USA' is at least in part responsible for the effect on the standard-of-living question, if people who believe that things are getting worse are prompted to think so and if that thought contributes to their answer about their own lives. The standard-of-living questions come after the health questions, which come after the ladder question and are strongly predictive of it. Earlier work has shown that the effects of context-setting questions can persist through many intervening questions (see, e.g., Bishop, 1987), which is again consistent with the idea that when people are confronted by a difficult question requiring cognitive effort or when they do not have a clear, readily available answer, they minimize effort by using recent relevant information—either the ladder itself or the 'way the USA is going' question. Such effects can persist long into the survey.

## 3. How do the answers matter?

The randomization allows us to compare those who did and did not answer the political questions, but it does not tell us what it is about the political questions that affects the subsequent answers or how the effect works. One of us, Deaton (2012), previously interpreted the political questions as exerting a negative effect on the mood of the respondents: that people should be irritated by being asked about politics was plausible given the deep unpopularity of Congress and politicians at the time. However, it is possible to provide some evidence on this conjecture because, for those who were asked the political questions, we know what their answers were and we can check whether their well-being scores were different depending on their answers. For example, do people who disapprove of President Obama's handling of his job have their ladder scores reduced more than those who approve? There is no separate experimental manipulation for addressing this question, so the analysis is essentially an ex-post subgroup analysis of a randomized controlled trial and has the usual disadvantages of such analyses. For example, support or opposition for President Obama is likely to be associated with other respondent characteristics, such as political affiliation, age, or race. Even so, we can compare the outcomes for those who did not get the political questions with those who did, and can separate the latter group into those who support or oppose President Obama, or who think the USA is or is not going in the right direction.

Table 3 shows, in the first column, the overall effect of asking the political questions; this is the coefficient on a dummy for the presence of the two questions in a regression with the ladder as the dependent variable. In the second column, there is a dummy for a negative response to the Obama question and a dummy for a negative response to the US question; it is the latter that matters, with the former exerting a small positive effect. The third

**Table 3.** The effects of answers to the political questions on the ladder: questions about opinion of President Obama and direction USA is going

| | | | | | | |
|---|---|---|---|---|---|---|
| Constant | 6.79 | | 6.80 | | 6.79 | |
| Political questions | −0.65 | (22.9) | – | | – | |
| Obama negative | | | 0.09 | (2.6) | – | |
| USA negative | | | −0.87 | (29.6) | – | |
| Obama OK, USA OK | | | | | 0.02 | (0.5) |
| Obama OK, USA negative | | | | | −0.86 | (28.6) |
| Obama negative, USA OK | | | | | 0.32 | (2.5) |
| Obama negative + USA negative | | | | | −0.78 | (17.6) |

*Notes*: Each column is a regression with the ladder as dependent variable. The first column repeats the information in Table 1, regressing the ladder on a dummy for whether the two political questions were asked. The second column 'splits' the political dummy into those who approve or disapprove of Obama and those who approve or disapprove of the direction of the country, while the last column 'splits' the political dummy into the four possible groups answering the two questions. Absolute *t*-values are in parentheses; standard errors clustered at the day level.

column has dummies for all four combinations of answers to the two questions. The negative effects are for being dissatisfied with the USA, and differ little whether they are or are not satisfied with President Obama. Perhaps surprisingly, there is a slight positive effect on the ladder from being negative about Obama's performance among those who are happy with the direction of the country

The lower scores on the ladder come from both thinking the country is going in the wrong direction and being asked to report the fact. As the control group was randomly selected, we can expect that, in both the control and treated group, 80% thought the country was going in the wrong direction. Thus, the lower scores on the ladder look like a context effect from asking this question. This is consistent with the argument that, when asked a difficult question to which they have no ready answer, they reach back in the 'stack' to find something that will serve as an answer (without awareness that they are doing so)—in this case the question about satisfaction about 'the way things are going in the USA'.

## 4. Context effects and intergroup comparisons

In this final empirical section, we turn to a different but related question, which is the extent to which context effects change the well-being rankings of different groups, particularly by age, sex, and race/ethnicity. While we recognize that, once again, the specific answers may be part of the mechanism—for example, that more blacks support President Obama, so that that question may elevate the well-being ranking of blacks—the specific answer effects are not the focus of these analyses.

Well-being measures are often used to make comparisons across groups, for example by gender, age, employment status, occupation, education, or place of residence. Such comparisons are arguably useful in policy, for example by making people aware of differences before they make choices, for incorporation into project evaluation, or for designing distributive policy. There are two separate issues here. The first, one of statistical significance, is the straightforward question of whether the context effects are significantly different

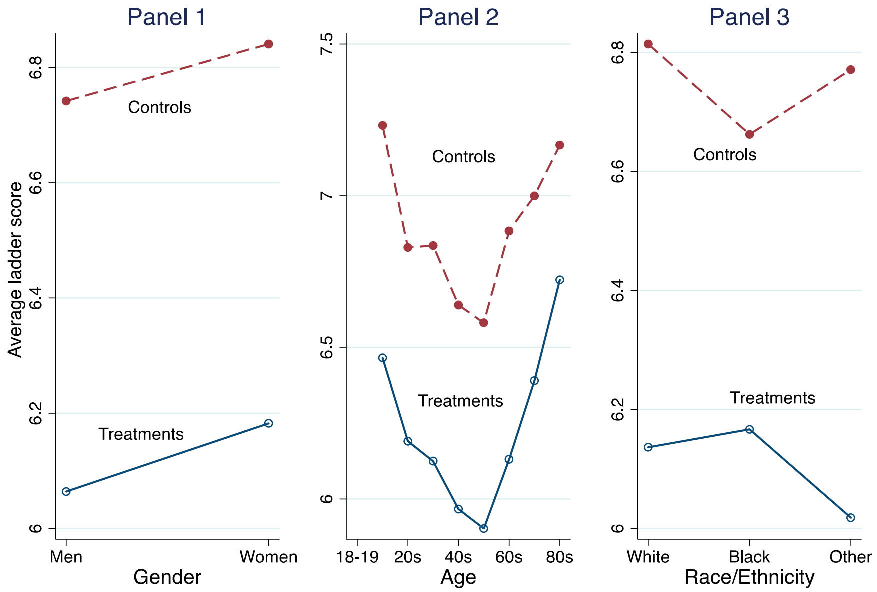

**Fig. 1.** Treatment effect of political questions, by gender, age group, and race/ethnicity.

across relevant groupings of the population. This can be analysed by a subgroup analysis of the randomized controlled trial in the Gallup data. The second question, about which it is more difficult to be precise, is whether the differences are large enough to matter for the kinds of comparisons that are usually made.

Figure 1 highlights some of the results. Panel 1 shows the standard result that women report higher ladder scores than men, by 0.10 of a rung in these calculations. The political question treatments reduce the ladder for both men and women. These numbers, although interesting, are small compared with the main effect for women of 0.10, so they are not close to being able to undercut the broader finding, that women rate their lives more highly than do men. Panel 2 plots the ladder by age group for those who were and were not asked the political questions. The graph shows the familiar U-shape of life evaluation with age. The *F*-statistic for the interaction of age group and treatment status is relatively low, and there is an almost uniform shift for all ages; the main effect of treatment is large and negative. It is hard to imagine a policy context in which such a small difference would matter, given that the U-shape is preserved. Panel 3 shows the ladder for whites, blacks, and all others (including those who refused to identify their race). The main story from this picture is that the ATE is quite different for blacks than for the other two groups. The negative treatment effect is markedly smaller for blacks.

We have drawn the corresponding figure for income, with results that are similar to those by age group, namely that the main effect predominates, with only minor changes across income groups (data not shown).

## 5. Discussion

We have shown that answers to questions about evaluative well-being are sensitive to the context in which the questions are asked, with the answers to previous political questions conditioning the answers to questions about well-being. The effects appear to be spillover effects generated by the answers to the political questions.

The question most affected by the political questions was the Cantril ladder question, which immediately followed, but the spillover effects did not evaporate quickly and affected the answers to related questions that were asked much later in the questionnaire. Of interest is the observation that a standard-of-living question, which like the Cantril ladder question may be difficult to answer, was affected, but not questions about less ambiguous content such as marital status, smoking, and race or ethnicity. This is consistent with idea expressed by Schwarz and others that difficult-to-answer questions are more susceptible to context effects (where questions about marital status and smoking are 'easy' questions). Furthermore, in analyses not shown, we observed that the hedonic questions about 'yesterday' used by Gallup, which may also be considered relatively concrete and easy to answer, are only slightly affected by the political questions. However, this interpretation is not entirely clear, because the hedonics are asked much later in the questionnaire, so two effects—carryover of context effects and ease of answering the hedonic questions—are confounded. Yet, the standard-of-living questions, which are also asked late in the questionnaire, are affected by the context questions, in spite of their position. Even so, we do not claim strong evidence from the current study about the relative sensitivity of hedonic and evaluative reports.

We also showed that context effects operate in different ways for different groups, for example for the elderly versus the young or for blacks relative to whites. Because the context effects likely work by content priming, by reminding people of their views about some perhaps loosely related topic (though we admittedly have no direct evidence of this), the effects will generally be heterogeneous, in part because the distribution of answers to the context question is different in different groups—such as blacks' versus whites' views on President Obama's performance—and in part because, even when they have the same answers, the effect on subsequent answers can differ across groups. We showed that these differential effects can be large enough to change rankings of well-being across policy-relevant socio-demographic groups. These observations may be important for evaluating group differences in well-being, because true group effects can be confounded with differential context effects, which could lead to erroneous conclusions. And unless the results of experiments of the types presented here are known, investigators may be entirely unaware of this potential threat to the validity of their studies.

We also note that a standard 'solution' for controlling context effects is to place the most important or potentially context-sensitive questions at the beginning of a questionnaire or interview, which does indeed eliminate any systematic context effects from preceding questions. The obvious issue with this solution, which is highlighted by this study's results with standard-of-living questions embedded later in the interview, is that only a single question can be first. Furthermore, questionnaire designers are not likely to know which of the many interview questions are the most potent in terms of producing context effects unless an unreasonable amount of pretesting is done. In this case, the interview could have protected the evaluative SWB question, but the hedonic questions would have remained at risk.

Our findings raise considerable difficulties for the uncritical use of evaluative well-being measures even when the context effects are similar for different groups, or even when they are identical for all individuals who have the same response to the context question. In particular, if we are interested in tracking well-being over time—for example, in the face of changing political events or economy-wide shocks over the business cycle—we need the changes to be changes in well-being, not changes in the balance of answers to a context question. The effects need to be real, not artefacts.

It may be argued that because evaluative well-being measures are reliable, at least in the sense of reliable used in the psychology literature (Diener *et al.*, 1999; Oishi *et al.*, 2003), they are therefore not susceptible to the contexts effects described here. Yet, the results presented here do indeed demonstrate a substantial and important context effect on an established evaluative well-being measure. The answer to this contradiction is that high test–retest or internal reliability coefficients do not in fact 'protect' measures from the kind of context effects with which we are concerned. To illustrate our central point, suppose that we are interested in monitoring national well-being over time, for example on a day-to-day basis, and that, again for illustration, we have a panel of individuals and the same people are interviewed every day. They are asked an evaluative well-being question and have been reporting a steady average of 7. We now introduce a question immediately before the well-being question that induces a context effect, a yes or no question about whether the country is going in the wrong direction. All of our panellists think that the country is in trouble and this reduces their own answer to the well-being question by a point. Those monitoring national well-being would be seriously misled, at least if they did not understand that what they were seeing is an artefactual context effect.

For purposes of argument, suppose that this reduction is the same for everyone, so that individual $h$ reports $S_h$ without the context effect and $S_h - 1$ with the context effect. For those interested in tracking, the national average drops from 7 to 6 overnight, which would be a serious matter for anyone who does not know what has happened. Yet, if we compare the two sets of responses before and after, $S_h$ and $S_h - 1$, and do a 'test–retest' comparison by calculating the correlation over people between the context-free and context-affected answers, we get a perfect correlation of 1, demonstrating that context effects can be independent of an instrument's reliability, at least when measured in this way.

We understand that this paper raises difficulties for much of well-being research as currently conducted and proposes no solution to those difficulties. Yet, it is better to be aware of the problems than to ignore them. And knowing about them is the first step to finding ways of neutralizing them and doing better.

## Acknowledgements

We are grateful to Gallup for granting us access to the Gallup–Healthways Well-being Index data. Both authors are consulting senior scientists with the Gallup Organization.

## Funding

We gratefully acknowledge financial support from Gallup and the National Institute on Aging through the National Bureau of Economic Research, Grants 5R01AG040629-02 and P01 AG05842-14, through the University of Southern California Grant 5R01AG042407-03, and through Princeton University's Roybal Center, Grant P30 AG024928.

## References

**Bertrand, M. and Mullainathan, S.** (2001) Do people mean what they say? Implications for subjective survey data, *American Economic Review*, 91, 67–72.

**Bishop, G.F.** (1987) Context effects on self-perceptions of interest 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, 179–99.

**Cantril, H.** (1965) *The Pattern of Human Concerns*, Rutgers University Press, New Brunswick, NJ.

**Deaton, A.** (2012) The financial crisis and the wellbeing of Americans, *Oxford Economic Papers*, 64, 1–26.

**Diener, E., Suh, E., Lucas, R., and Smith, H.L.** (1999) Subjective well-being: three decades of progress—1967 to 1997, *Psychological Bulletin*, 125, 276–302.

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

**OECD** (2013) *OECD Guidelines for Measuring Subjective Wellbeing*, OECD, Paris.

**Oishi, S, Schimmack, U., and Colcombe, S.J.** (2003) The contextual and systematic nature of life satisfaction judgments, *Journal of Experimental Social Psychology*, 39, 232–47.

**Schwarz, G.** (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, 461–4.

**Schwarz, N. and Strack, F.** (1999) Reports of subjective well-being: judgmental processes and their methodological implications, in D. Kahneman, E. Diener, and N. Schwarz (eds) *Well-Being: The Foundations of Hedonic Psychology*, Russell Sage, New York, 61–84.

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