

NBER WORKING PAPER SERIES

MEASURING SELF-REPORTED WELL-BEING

Ori Heffetz
Yehonatan Caspi

Working Paper 33847
<http://www.nber.org/papers/w33847>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2025

We are grateful to Dan Benjamin for early comments that significantly improved the manuscript; to Angus Deaton, Marc Fleurbaey, and Alex Rees-Jones for additional valuable comments; to Luca Braghieri, Pablo Brañas-Garza, Lina Diaz, Sarah Eichmeyer, Björn Frank, Elena Fumagalli, Laura Fumagalli, Matthew Gentzkow, Thomas Graeber, Johannes Haushofer, Daniel Houser, John Ifcher, Doerte Junghaenel, Richard Lucas, Mike McBride, Vincent Pons, Christopher Roth, Stefan Schneider, Lena Song, and Marta Walentynowicz for confirming the accuracy of our coverage of their work (of course, any remaining errors are our own responsibility); and to Keren-Or Barashi Gor and Ruth Zussman for research assistance. For financial support, we are grateful to NIH/NIA grant R01-AG065364 to Hebrew University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, other funding bodies, or the National Bureau of Economic Research. The authors have no material financial interests that relate to the research described in this paper.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2025 by Ori Heffetz and Yehonatan Caspi. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Measuring Self-Reported Well-Being
Ori Heffetz and Yehonatan Caspi
NBER Working Paper No. 33847
May 2025
JEL No. C83, C90, D90, I31

ABSTRACT

What should researchers consider when designing experiments that (also) collect self-reported well-being (SWB) data? Focusing on experiments in economics, we examine the motivation behind SWB-data collection, survey leading past examples, and highlight potential pitfalls and their proposed countermeasures.

We offer three main messages and a call to action. First, SWB measures should be handled with caution, especially in experiments. Second, despite their limitations, SWB measures can be used cleverly in the lab to provide evidence on questions that choice data alone cannot answer. Third, when collected, analyzed, and interpreted with appropriate caution, SWB measures can be important policy-evaluation outcomes, complementing the inherently incomplete picture provided by more traditional outcomes.

We call on researchers to carefully and thoughtfully collect (a variety of) SWB measures in their online, lab, and field experiments. Such a joint, decentralized effort would also mean that over time, SWB data get explored, accumulated, and, hopefully, better understood.

Ori Heffetz
Cornell University
Johnson Graduate School of Management
and The Hebrew University of Jerusalem
and also NBER
oh33@cornell.edu

Yehonatan Caspi
The Hebrew University of Jerusalem
yehonata.caspi@mail.huji.ac.il

1 Introduction: A Tale of Three Studies

In standard social-science experiments, individuals are randomly assigned into those who undergo some treatment and those who do not (a control group), and an important outcome is eventually measured on everybody. Now imagine a “dream” experiment: The treatment takes just a few seconds to administer and is otherwise costless—and its measured effect is found to be considerable. The treatment effect is comparable, for example, to the average difference in the measured outcome between the employed and the unemployed, or between those reporting annual income \$35,000–\$49,999 and those reporting \$50,000–\$74,999.

What is the important outcome and what is the magic treatment? As reported by Deaton (2012), the outcome is self-reported well-being (SWB)—the answer to a survey question about a respondent’s evaluation of their life. Specifically, a Cantril ladder (Cantril, 1965):

Please imagine a ladder with steps numbered from zero 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. On which step of the ladder would you say you personally feel you stand at this time?

Such outcomes are increasingly analyzed by economists, who have been slowly lifting a traditional taboo on non-choice evidence. The magic treatment, unfortunately, turns out to be a context effect: another survey question that is asked prior to the SWB question, or a short explanation on how to think about the SWB question. A first main message of the present chapter: SWB measures should be handled with caution, especially when used in experiments.

Now imagine a completely different scenario. You are interested in prosocial behavior, and want to answer the question: What makes lab participants better off: receiving from the experimenter €100 or having the experimenter instead donate €350 to save a stranger’s life? Having participants choose between the two options may not answer your question because while they may be better off *receiving* €100, they may be worse off *actively choosing* to take the money over having a human’s life saved (Kőszegi and Rabin, 2008).¹ Enter SWB data: Falk and Graeber (2020) find that participants in €100-received-versus-€350-donated lotteries whose (exogenous, randomly assigned) outcome is receiving the money do indeed report higher SWB a month later (while possibly reporting lower SWB at the end of the lab

¹More generally, as Kőszegi and Rabin (2008) observe about less dramatic choices that economics-lab participants often face (p. 1822), “if a person dislikes being allocated less money than somebody else but dislikes even more intensely hurting others, we will never observe in her behavior that she dislikes coming out behind.”

session). Importantly, this result holds also for participants whose (incentivized) behavior in the lab indicates a preference for (actively choosing) the *other*, life-saving option.² A second main message of this chapter: In spite of their limitations, SWB measures can be used cleverly in the lab to provide evidence on questions that choice data alone cannot answer.

As a third tale, imagine a field experiment evaluating the effects of unconditional cash transfers on poor households in a developing country. Haushofer and Shapiro (2016) conduct such an experiment, and collect a long list of outcome variables. These also include well-being measures: a physiological indicator—the stress hormone cortisol—and a long list of survey-based, SWB questions. The authors find consistent effects on SWB, including on self-reported stress, but no main effect on cortisol. Could this divergence reflect experimenter demand effects on SWB? The authors are well aware of this possibility, take countermeasures to mitigate it, provide some reassuring evidence, and cautiously conclude: “In the absence of a gold standard measure that assesses well-being directly, it is not clear which variable should be given priority.” A third main message of this chapter: When collected, analyzed, and interpreted with appropriate caution, SWB measures can be important policy-evaluation outcomes, complementing the inherently incomplete picture provided by more traditional outcomes alone.

Context effects, demand effects, and other potentially unintended consequences of a study’s implementation details are routine worries for social scientists. However, SWB responses may be subject to such worries more than other outcomes whose measurement is covered in other chapters of this Handbook, such as measures of risk, time, and social preferences, effort provision, willingness to pay, discrimination, dishonesty, and physiological responses. One potential reason is that SWB questions are never monetarily incentivized. Another is that they can be particularly taxing, emotionally or cognitively. Because they ask people to do what many think is impossible—namely, to reduce (aspects of) their multifaceted lives into a rating on a scorecard—respondents may look for more mental shortcuts when answering them. A third potential reason for worry is that a vague, general question about one’s life or one’s emotions, elicited by an experimenter who has just treated an individual, could come across as intending to ask about something else. A specific main message of this chapter, exemplified especially by the first, cautionary tale above: Social-science experimental treatments may have large effects on SWB *for the wrong reasons*.

²Recently, Bernheim et al. (2024) propose to address the problem highlighted by Kőszegi and Rabin (2008) using an approach that shares a main design element in Falk and Graeber (2020). As they put it in their abstract: “The standard revealed-preference approach to welfare economics encounters fundamental difficulties when the act of choosing directly affects welfare ... We address this problem by developing an approach that ... measures welfare by blending choice-based methods with self-reported well-being techniques.”

At the same time, when handled thoughtfully, SWB measures are readily available and cost-effective outcomes that can yield important insights. Indeed, one would be hard-pressed to make an argument *against* routinely collecting them in experiments. Adding a single-question SWB measure, or a small set of questions, when relevant, to a lab experiment, do not significantly add to its duration and cost. Adding a more substantial battery of measures to a field study takes more effort and is generally more costly—but its costs may still be negligible relative to other fixed costs of running the study. A specific main message of the chapter, exemplified by the second and third stories above: If treated with appropriate care and thought, SWB measures could help assess the *intentional* (“real”) effects of experimental treatments on (aspects of) well-being.

The goal of this chapter is to encourage researchers to carefully and thoughtfully collect SWB data in their experiments. Even when we do not have all the answers—for example, when context and demand effects cannot be completely ruled out—such data collection has an additional, positive-externality benefit: The more SWB data are collected, the better researchers can study, understand, and improve both the measures and their use in the future.

The chapter proceeds as follows. In Section 2 we briefly discuss the SWB measures, distinguish between their use in observational versus experimental studies, and discuss general methodology considerations that are analyzed in more detail in later sections. While most past reviews focus on observational studies, our focus here is experiments. We also ask: “What do SWB measures capture?” We propose to place researchers’ (expressed or implied) views on this question on a continuum that stretches between two extremes: a “Means to an end” view (the end is some underlying well-being concept) versus a “Just the question” view (the main interest is the SWB question itself). We end the section with a brief discussion of the types of assumptions that are necessary in support of each view.

In Section 3, we review several examples of past experiments that use SWB measures as outcomes. We focus on those run by economists. Unlike most reviews, however, we do not focus on their findings. Rather, we use them to demonstrate (a) the range of research questions, experimental treatments, and SWB measures that past research has studied; (b) the motivation for using SWB, and specific SWB measures, as communicated in past research; (c) how past research discusses, and attempts to deal with, issues such as demand effects; and (d) what insights can be drawn when combining said past research with research on SWB methodology in economics. These past studies can generally be divided into two groups. A first group of experiments, often conducted in the lab and online, aims to explore economic theories, assumptions, and hypotheses (e.g., regarding social preferences). Such studies are most naturally interpreted as using SWB as a utility proxy, although, as we

document below, they are often explicitly motivated by other, or by less clearly specified, notions of well-being. A second group, conducted in the field, studies the effects of different policy interventions (e.g., unconditional cash transfers). These studies use SWB as merely one available assessment metric, often collected in program-evaluation studies alongside more traditional outcomes like income, consumption, and health. As we document below, their authors’ motivation for using SWB varies.

Both groups of experiments are mostly latecomers into a large and growing—and originally non-experimental—literature that uses SWB data from large-scale surveys such the General Social Survey (GSS) to answer economic questions that are difficult or impossible to answer with choice data alone. Naturally, that broader literature raises the question: Under what conditions is an SWB measure justified as a utility proxy?

Since a recent review, titled “From Happiness Data to Economic Conclusions” (Benjamin et al., 2024), takes up this question directly, the present chapter does not cover it. That review makes a distinction that we adopt here, “between the empirical measures—the SWB data—and the underlying theoretical and philosophical concept they aim to measure—the well-being (or welfare, or utility) of certain individuals during a certain period of time” (p. 360). The review offers an underlying theoretical economic framework, discusses the potential advantages of SWB over choice data, and lays out a list of necessary assumptions for SWB to be informative about preferences. Those range from the assumption that SWB is ordinally equivalent to lifetime, forward-looking, or flow utility—a sufficient assumption for using SWB as a measure of an individual’s ordinal preferences, but far from sufficient for applications that, e.g., compare mean SWB across groups of individuals—to additional, stronger assumptions involving cardinality, interpersonal comparisons, and group well-being. The latter include the assumption that SWB is interpersonally comparable in both levels and differences; that people use the SWB response scale in the same way; and that the scale is such that mean SWB is a valid social welfare function (see also Adler (2019), e.g., Chapter 2). Finally, that review also discusses survey-methodology assumptions that may be less relevant for experimental work, e.g., because they either relate to issues that are addressed, at least partially, by the random assignment inherent in experiments or, like nonresponse bias in large surveys, take a different form in experiments.

This chapter therefore focuses on experiments’ *other* potential uses of SWB. Rather than specifically as a utility proxy, some researchers appear to have a more general (or vague) interest in SWB that may or may not be supported by some (explicit) underlying view of well-being. (In the extreme—although we are aware of no specific examples—researchers could in principle be interested in SWB purely for what it is: experimental participants’ answers to questions about their feelings and life evaluations.) By focusing on this more “inclusive”

view on what these answers may reflect, the present chapter can focus on experimental-design considerations that are relevant to a broader set of researchers. These include researchers who do not commit to the traditional economist’s requirement, namely, that SWB is informative only to the extent that it informs us about the satisfaction of preferences. As we demonstrate in Section 3, generously quoting from the original studies, these researchers indeed include those who conduct field, policy-relevant studies—the second group of experiments mentioned above. But as the quotes below also suggest, even researchers in the first group above, who explicitly test economic assumptions or hypotheses in the lab (e.g., that well-being is related to relative income), often seem to be interested at least in part, or even mainly, in some other notion of well-being. Accordingly, we focus in this chapter on a minimalist main question: Under what types of experimental settings is an SWB measure more likely to reflect *some* notion of participants’ underlying well-being—rather than, unintentionally, something else?

In Sections 4 and 5 we therefore return to the cautionary tale above and review evidence on potential *unintentional* experimental treatment effects on SWB measures. We focus on context effects (in Section 4) and demand effects (Section 5). Both have long been studied by psychologists and other social scientists, with different researchers at different times drawing different conclusions and emphasizing different lessons learned. We sift through some of this rich and sometimes mixed evidence, aiming to draw our own conclusions and—importantly—offer advice based on an updated synthesis.

Our advice in Sections 4 and 5 is guided by three considerations: It should, whenever possible, be based on (a) the most recent empirical evidence (with its bigger samples and other improved research methods); (b) evidence that we believe is the most relevant to SWB (and other self-reported) measures; and (c) evidence that seems the most relevant in the particular setting of experimental studies. We aim at giving *practical* advice. That some of it is expressly tentative—rather than an authoritative Do and Do Not list—reflects our impression of the state of evidence and knowledge. We believe that more research is needed before one can write a comprehensive How To guide on measuring SWB in experiments. That research will be facilitated by more SWB data collection—which should not wait, in a vicious circle, for a How To guide. Moreover, existing evidence already highlights potential pitfalls one needs to think about when designing experiments and, in several cases, already suggests ways to help mitigate undesirable effects. When we believe it is justified, we try to provide readily implementable advice.

We conclude in Section 6. Following past work, we provide a table that summarizes the advice that we give throughout the chapter. We view the table as a go-to, yet work-in-progress aid: We hope that researchers both find it helpful and, in the future, help to improve it.

2 SWB in Experiments

2.1 The SWB Measures

SWB measures are often categorized into three or four types. Evaluative or cognitive measures, such as the ladder question from the introduction, ask respondents for a (cognitive) evaluation of their lives. Hedonic or affective measures—sometimes referred to as experiential measures—ask about specific emotions felt (or experienced) during a specific point in time; we provide examples in the next section. They are often further subdivided into positive affect, e.g., happiness or joy, and negative affect, e.g., anxiety or sadness. Finally, eudaimonic measures include questions related to having a meaningful, purposeful, or worthwhile life.

Naturally, lab experiments do (and should) focus on affective measures—it would be troubling if winning a small amount of money in the lab changed one’s life evaluation or meaning. In contrast, policy-evaluation RCTs, that sometimes involve potentially life-changing field interventions, should not (and do not) rule out any of the above SWB-measure categories. Indeed, some such RCTs, reviewed in the next section, elicit a combination of several measures.

SWB measures are also referred to as eliciting *general* well-being, e.g., the ladder question, versus *domain-specific* well-being, e.g., satisfaction with one’s employment, finances, or relationships. Job satisfaction, for example, has been studied by psychologists for almost a century, with economists joining in more recently (see, e.g., Card et al., 2012, for an experiment studying the effect of peer salaries on job satisfaction.) Some of the studies reviewed in the next section measure both general and domain-specific SWB; indeed, question wording is sometimes rather similar across the two, with only a one- or two-word difference specifying the domain. Section 7.2 in Benjamin et al. (2024), and the references therein, give advice on developing multidimensional utility proxies by first measuring single domains and/or aspects and then estimating weights to aggregate them into an individual-level general well-being index. We hope these methods will eventually become a gold standard. However, they are data-hungry, and hence may be less practical in experiments with small-to-midsized samples. We do not discuss them here and, more generally, we do not focus on domain-specific measures in this chapter.

Past reviews of SWB methodology discuss important considerations for SWB researchers before collecting, analyzing, and interpreting SWB data. Due to random assignment, many of these do not apply, or apply less, to experiments. For example, Pavot (2018) mentions systematic cultural differences, including language-related differences, in SWB-question interpretation or in response-scale use; important as they are in cross-group comparisons, these are rarely relevant in experiments. Similarly, differences across age groups (i.e., cohort

effects), e.g. due to different severe geopolitical or economic experiences, such as wars or financial crises while growing up, are all but irrelevant. Benjamin et al. (2024) observe that potentially non-representative study samples, due to low and decreasing response rates in commonly used social surveys, make it increasingly difficult to draw conclusions regarding target populations; these concerns take a rather different form in experiments.³

At the same time, typical lab and field experiments that collect SWB data may raise other concerns that are either less severe or entirely nonexistent in studies that analyze SWB data from large-scale surveys. One increased concern is larger measurement error due to smaller sample sizes. It makes repeated measurement, of a variety of outcomes, and using multi-item measures especially useful in experiments. We generally recommend such multifaceted measurement throughout the chapter; we specifically document it, along with sample sizes, in the SWB experiments we review in Section 3.

In Sections 4 and 5 we focus, respectively, on what we suspect are two of the biggest concerns in experiments: context effects and demand effects. We approach them as cases of the general concern that the experimental manipulation may unintentionally cause differences, across treatment groups, in how respondents interpret and react to SWB questions.⁴ This concern, unique to experiments, replaces the concern above that such differences may preexist across respondent groups in non-experimental data. In particular, Section 4 emphasizes that an experimental manipulation is always, by definition, also a manipulation of context. An experimental effect on SWB could therefore also be, at least partly, an unintentional context effect. Specifically, the same SWB question asked in different treatments—i.e., in different contexts—may, unintentionally, be interpreted rather differently, as participants use the context in which it is asked as a cue. An SWB question may therefore elicit different responses across different experimental conditions *for the wrong reasons*.

Section 5 in turn highlights that in contrast with SWB questions that are asked as part of routine large-scale surveys and are then used in types of analysis that are largely unforeseen by respondents, SWB questions in experiments are often asked by the experimenter, within the experiment, after participants in some treatment groups received something of

³Low response rates are an increasing concern in large-scale surveys. (Recent response rates in the GSS, for example, are around 50 percent, and they are much lower in many other large surveys.) If respondents with different demographics differentially self-select into survey participation, SWB correlations that are observed in the data may fail to hold among nonrespondents and, hence, may fail to generalize to the underlying population (Heffetz & Rabin, 2013). But in experiments, the main studied effects are causal; the issue then becomes that of external validity, or coverage, rather than that of self-selection on the independent variable—assuming, of course, no incompleteness (in the lab) or attrition (in the field) that are conditional on treatment. We return to these points below.

⁴Another potential case of this general concern is that the experimental manipulation may cause differences across treatment groups in SWB scale use (Benjamin, Cooper, Heffetz, et al., 2023). As we suspect that this concern is secondary in our context, we do not focus on it in this chapter.

value—sometimes, directly from the experimenter. In such settings, participants may easily infer the researcher’s hypothesis or goals, and may respond by adjusting their SWB reports accordingly. In contrast with context effects (Section 4), which past research on SWB rarely explicitly considers, researchers are, for the most part, fully aware of potential demand effects. Indeed, as we discuss in detail in Section 3, researchers commonly consider them, and attempt to both reduce them *ex ante* and assess them *ex post*.

Research on SWB methodology is a vibrant field, in both economics and psychology. SWB measures are not only increasingly *used* by general-purpose economists; they are also increasingly *studied* by “happiness” economists. This chapter is written for the former group. We aim to provide relevant information and, occasionally (when we feel it is justified), guidance to practitioners whose purpose is to use SWB data, rather than to study them. Moreover, as the example research we review in Section 3 highlights, we focus on research motivated by either economic or policy-oriented questions. We do not cover SWB use in research on therapies, or on psychological interventions aimed at increasing all sorts of emotional (and other) well-being, from physical activity to stress-management interventions. For a taste of these, see the different contributions in the Interventions part (pp. 700–754) in Diener et al. (2018).

2.2 What do SWB Measures Capture?

Researchers who consider adding SWB measures to their experiment should first ask themselves why: What are they trying to measure? Views on this question in existing experimental work seem to vary (as we demonstrate in the next section, providing direct quotes as much as possible). Most researchers appear to motivate their use of SWB data with a view that lies somewhere between the following two extremes.

On one extreme, a researcher may view their SWB measures as a means to an end: they are interested in SWB responses only because, and to the extent that, they provide information about some underlying—and, hopefully, normatively appealing—concept of well-being. Some economists (e.g., Layard, 2005), and many psychologists, may have roots in a hedonistic theory of well-being (Bentham, 1789, Kahneman et al., 1997); they may hence hope to capture reliable inputs into the calculus of a respondent’s pleasure and pain. Most economists, in the tradition of neoclassical welfare economics, may have a preference-satisfaction view; they may hope to capture reliable information for learning about preferences. Other economists (most famously, Sen, 1985) and some philosophers (Nussbaum, 2000) have an objective-list view; they may hope to capture information about the necessary items on the list (e.g., capabilities and freedoms).

From this “Means to an end” point of view, SWB responses have to satisfy certain assumptions. The specific assumptions depend on the researchers’ specific view, goals, and analysis of the data. Consider the preference-satisfaction view, that is most relevant to economists. At the very least, SWB responses have to be ordinally equivalent to a fixed utility function that represents stable preferences, for a given individual across different experimental conditions and over time. Then, in within-subject experiments, the direction of change in each respondent’s SWB across conditions is informative regarding the direction of change in their well-being; a researcher can then report, for example, the share of respondents whose well-being improves or worsens under a certain condition.

Even this minimalist assumption, however, may not hold when respondents use the SWB response scale differently over time. Moreover, including in the analysis statistics that require cardinality (e.g., mean difference across conditions) or that are based on cross-group comparisons (e.g., when using a between-subjects experimental design)—both common practices in almost all experiments—require much stronger assumptions, as discussed in great detail in Benjamin et al. (2024). As also discussed there, existing evidence does not always support these assumptions, and random assignment may address only some of them, only to some extent, and only under other strong assumptions. (For example, only if the treatment does not affect scale use and is not correlated with individual scale-use parameters; see also Benjamin, Cooper, Heffetz, et al., 2023.) Finally, that paper also reviews evidence on the question: Which types of SWB measures—evaluative or hedonic—better capture preferences? Evaluative questions are general, yet often vaguely phrased. For example, they often do not explicitly state the social circle (e.g., self versus family) or the time period (e.g., today versus this year) they ask about (Benjamin, Debnam Guzman, et al., 2023). In contrast, hedonic questions are simpler, more explicitly focused, and easier to interpret. However, evidence suggests that it is the former that get closer to capturing preferences, at least under some circumstances (Benjamin, Heffetz, Kimball, & Rees-Jones, 2014).

On the other extreme—and in sharp contrast to the “Means to an end” view—is what may be called the “Just the question” view. It is anchored in the observation that vague questions such as “How is your family doing these days?” are prevalent in everyday conversation. A social scientist may therefore be curious as to how responses may change under different experimental conditions. From this point of view, first, the question does not have to be well defined: the person asking the question in everyday life may themselves not have a specific notion in mind regarding what they mean by “your family” (the exact social circle to be included in the question) or by “these days” (the exact time frame the question refers to). Second, the question does not have to measure an underlying, philosophically defensible notion of well-being: again, the asking person may themselves not have a particular

underlying well-being concept in mind. Third, the question does not have to mean the same thing (i.e., be interpreted in the same way) to different people, or even to the same person at different times or across contexts—including across experimental conditions! And it is of course perfectly fine if different people focus on different aspects of their lives when answering it (e.g., because they have different preferences) or even use words differently when responding to it (e.g., if one’s “Very good indeed!,” or “10/10,” convey the same evaluation as another’s “Fine, I guess,” or “8/10”).

For pure “Just the question” researchers, no assumptions are needed regarding what SWB data capture. Indeed, even context effects and demand effects are not an issue—they are merely factors that affect answers. We are not aware of specific such purists, however. While some researchers’ vague statements may suggest that they may not have a clear idea of their own position on this issue, we have not encountered an explicit “Just the question” motivation in any of the studies below. We would find such a motivation difficult to defend normatively. We mention this extreme view here mainly to draw the full spectrum of logical possibilities.

The rest of this chapter is aimed at those who are somewhere in between: They are interested in SWB because they are interested in underlying well-being, but they are not committed to a single definition of well-being. Such researchers may hesitate between, for example, a hedonistic view, thinking that happiness and pain are everything; an objective-list view, thinking that happiness and pain are items on the list; and a preference-satisfaction view, thinking that they are inputs into preferences. Some such researchers may wish to intentionally maintain this normative flexibility. Or they may simply not be sure what (underlying) well-being is. But whatever it is, no such researchers are likely to expect it to be subject to extreme context and demand effects. The following sections, which focus on such minimalist (loosely defined) conditions, are written for them. We do reiterate here that economists who do want to maintain consistency with a traditional preference-satisfaction view of well-being should consult Benjamin et al. (2024). Such economists should view the following sections as discussing important considerations, rather than sufficient conditions, for collecting SWB measures informative about underlying well-being.

3 SWB in Past Economics Experiments

We now discuss a few examples of past economics experiments using SWB measures. Our purpose is to demonstrate the types of questions researchers ask, SWB measures they use to answer them and, generously using direct quotes, motivation they give for using these measures and design approaches they take for tackling potential issues. We say nothing about

non-experimental work in economics—a half century of research, starting with Easterlin (1974), that utilizes SWB data collected as part of large-scale social surveys;⁵ we say nothing about experiments in other disciplines—importantly, psychology—using SWB data;⁶ and even within experimental work in economics—the focus of this section—we do not aim to provide an exhaustive survey. Rather, we focus on example research that helps make our advice in later sections more concrete.

One important class of economics experiments that we mostly leave out are those that use SWB as an *independent* variable. Ifcher et al. (2022), in a review dedicated to SWB in online and lab experiments, cover such studies in depth. Typically using a mood-inducement procedure involving, e.g., a video clip or a gift, such studies investigate the procedure’s effect on different (non-SWB) dependent variables. For example, Ifcher and Zarghamee (2011) explore whether positive affect influences time preferences. Participants watch a brief video, either neutral (control group) or designed to induce mild positive affect (treatment), both selected from prior research in psychology, and then make incentivized choices between money today and money in the future (using the Becker-DeGroot-Marschak (BDM) mechanism). For our purposes, an important design feature of this particular study is its thoughtful inclusion of what we view as a “placebo SWB measure”: While the authors’ experimental manipulation is designed to induce *mood*, their manipulation tests include both a measure of mood—which should show an effect—and, importantly, a measure of what they refer to as long-term happiness—which should not.⁷ In a later section, we recommend such use; we have not seen it widely discussed outside this study.⁸

⁵The review by Benjamin et al. (2024) mentioned above lists on its first page around twenty excellent earlier reviews of the SWB literature in economics. We refer to some of them below.

⁶Many chapters in the *Handbook of Well-Being* (Diener et al., 2018) review the SWB literature in psychology. We discuss some of them below.

⁷Ifcher and Zarghamee (2011) measure mood using the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), where participants rate the extent of seven positive and nine negative affect items on a scale from 1 to 10 (amusement, arousal, contentment, happiness, interest, relief, and surprise; anger, confusion, contempt, disgust, embarrassment, fear, pain, sadness, and tension). They measure happiness using the question:

Taken all together, how would you say things are these days—would you say that you are...

Possible responses range from 1 (completely unhappy) to 7 (completely happy). The authors motivate this happiness question as a placebo question (p. 3115):

This measure is similar to the ones used in the GSS and the World Values Survey (WVS), each of which has been used extensively in the happiness-economics literature as a measure of long-term happiness. . . . This question was asked after the mood-inducement procedure and the time-preference questions, but because of its broad scope, it should not have been affected by the treatment (Kahneman et al., 1999).

⁸While we do not report here the specific findings in Ifcher and Zarghamee, 2011, we note for completeness

3.1 Exploring Economic Theories, Assumptions, or Hypotheses

Reference dependence and (especially) social aspects of well-being loom large in past SWB lab experiments by economists. One reason could be the twin difficulties of identifying causality and using choice in these contexts. We briefly review five example studies below; their authors are interested in topics that revolve around aspirations/expectations and social comparisons, reference points and the reference group, relative income, social preferences, and prosocial behavior.

In our first example, McBride (2010) notes that recent research questions economists’ fundamental beliefs that “more is better” and that “economic growth should be a primary goal of public policies.” Motivated by studies that associate higher income with higher SWB but also with higher consumption aspirations that, over time, could lower SWB, the author sets out to study how aspirations form and adapt over time and how they affect happiness. He runs a lab experiment investigating the impact of past payments, social comparisons, and expectations on reported SWB. Participants ($N \approx 100$) play 25 rounds of matching-pennies games with a computer, who adopts different “types” differing in heads-versus-tails play probabilities. In each round, participants know their opponent’s type prior to playing, and learn the round’s outcome (and their payment) after playing. Depending on treatment, they may also learn all other participants’ outcomes for the round, on average or by opponent type. They are then asked:

How satisfied are you with the result of this round?

The response scale ranges from 1 to 7, with 1, 4, and 7 labeled “very dissatisfied,” “satisfied,” and “very satisfied,” respectively. Using this measure as his main outcome, the author finds negative effects of both expected earnings (reverse-engineered from the computer’s type) and others’ earnings, no effect of past earnings (from previous rounds), and a positive effect of current earnings. We note that the SWB question is very clear about what it asks, which may reduce context effects (see Section 4). At the same time, rather than asking about current mood or emotion in general, the SWB question asks about satisfaction with a specific experimental result—which may increase demand effects (see Section 5; while the author does not discuss potential demand effects, his findings of an effect of current earnings but not of past earnings may weaken such concerns). We also note that the author talks about happiness and satisfaction interchangeably—a recurring theme in this section, as well

that Camerer et al. (2016) fail to replicate them. However, they fail to replicate even the mood inducement (i.e., the original, statistically weak, effect of the manipulation on mood), possibly due to participants in the later replication study no longer finding funny the original study’s video, whose actor had since famously committed suicide.

as in the broader literature.⁹

In our second example, Ifcher et al. (2020) are motivated by similarly big, related questions: “Does being richer than others make people happy? And if so, why?” Referencing John Stuart Mill, disputedly credited with the claim that “men do not desire merely to be rich, but richer than other men,” the authors run an online experiment ($N \approx 900$) to better understand social comparisons and the relative income effect. They measure participants’ SWB before and after an external shock that reveals (a) the number of experimental points the participant and another participant received, and (b) whether these points have monetary value. Their experiment is designed to address identification issues in past work that studies these questions using large-survey SWB data (and that finds a negative relationship between others’ income and one’s own SWB). In their experiment, the authors find, among other things, that participants report significantly greater SWB when receiving points with monetary value than without it, but equally reduced SWB when receiving fewer monetized versus non-monetized points than others—suggesting that while money has absolute, non-symbolic value, its relative value may be purely symbolic.

Their SWB measure is a Mood Short Form (MSF; Peterson and Sauber, 1983) score. Ranging from 4 to 20, it adds together responses on a five-point Likert scale (“Strongly disagree” = 1, “Disagree” = 2, “Neither agree nor disagree” = 3, “Agree” = 4, “Strongly agree” = 5) to the following four items (the last two are reverse-coded):

Currently I am in a good mood.

As I answer these questions, I feel very cheerful.

For some reason I am not very comfortable.

At this moment I feel ‘edgy’ or irritable.

Item order was randomized by participant, and to mitigate participants’ inclination to anchor post- to pre-revelation reports, elements from the Basic Psychological Need Satisfaction Scale (BPNSS) were incorporated into both surveys.

The authors’ choice of the MSF and BPNSS was based on an earlier “discovery study,” whose purpose was “to determine SWB measures’ sensitivity to money-rewards of the size paid in experiments.” As the authors explain (p. 1211):

⁹As McBride (2010) observes in his footnote 13, much of the income-happiness literature uses the terms happiness and satisfaction interchangeably and views them as empirically equivalent. Like others, he too is not explicit about what underlying well-being concept, if any, he hopes to capture. Rather, he too motivates his use of SWB by referring to psychologists’ conventions and using only very general terms (talking about capturing “relevant aspects” of well-being; p. 266): “The form of this question matches the convention used in happiness surveys (Schwarz & Strack, 1999), and although answers to these subjective questions suffer from various imperfections, a widespread conclusion is that such data meaningfully capture relevant aspects of happiness or satisfaction.”

For the discovery study, we compiled a comprehensive list of SWB items and scales that had been previously used in the SWB literature, including eudemonic, evaluative, and experiential measures; these came from oft-cited SWB datasets (e.g., the World Values Survey), and contained popular SWB measures (e.g., Cantril’s Ladder, a single-item measure) and scales (e.g., Positive Affect Negative Affect Scale, a multi-item measure). In the discovery sample, subjects completed a total of 92 SWB items (8 single-item and 7 multi-item measures), and, of these, the MSF was found to be the most responsive to revelation and the BPNSS among the least responsive.

While the authors’ rationale for running an early discovery study is well motivated and transparently explained, their findings also raise questions regarding why some SWB measures are found more sensitive to their experimental treatments than others; whether the effects they find are SWB-measure-specific; and the extent to which they generalize. For example, as is standard practice in economics, the authors use words such as “happiness” and “mood” loosely and interchangeably when describing their findings; but the PANAS is no less a measure of mood and happiness than the MSF. The authors used it in previous work (see footnote 7) but, based on the discovery study, not in this work. Would it make a difference? We return to the question of whether and when the measure matters below.

In the paper’s last paragraph, the authors appropriately discuss the caution required when trying to infer preferences from SWB. They explicitly allow for both the possibility that SWB change is, and is not, compatible with revealed preferences (p. 1229):

Our approach, of course, begs the question of whether SWB-change is predictive of choice, and whether a revealed-preference approach would result in the same patterns we observe. In Benjamin, Heffetz, Kimball, and Rees-Jones (2014) ... [p]redicted SWB is found to be the best, but not the only important, predictor of actual preference-rankings. ... Further research is necessary to directly address whether our mood-change approach is compatible with a revealed-preference approach. If so, it could substitute for revealed-preference in other contexts, like identifying preferences for earned versus unearned income, or estimating parameters related to loss aversion, time-, risk-, or competitive-preferences. If not, it could potentially complement revealed-preference.

In our third example, motivated by closely related questions regarding social comparisons, Fumagalli and Fumagalli (2022) aim to study the impact on SWB of the gender composition of the (perceived) reference group. Hypothesizing that women, who “do (or perceive they do) less well than men in several domains,” would report higher SWB when they compare

themselves with other women, they report the results of a survey experiment ($N \approx 1,000$) in which treatment (wo)men respondents are asked (as rephrased, reformatted, and reported by the authors):

How dissatisfied or satisfied are you with your [domain] if you compare yourself with other [(wo)men]?

and control respondents are asked:

How dissatisfied or satisfied are you with your [domain]?

Answers are recorded using Likert scales with the options: “completely dissatisfied,” “mostly dissatisfied,” “somewhat dissatisfied,” “neither satisfied nor dissatisfied,” “somewhat satisfied,” “mostly satisfied,” and “completely satisfied.” The survey includes a set of four questions, with [domain] replaced with “health, household income, the amount of leisure time, and life overall.” It is run as part of the UK Understanding Society Innovation Panel; the control question is the Understanding Society standard SWB question.

The authors find that in some domains, treatment women report higher satisfaction than control women. No effects are found on men. Notice that the control question does not specify a comparison group. While the authors’ preferred interpretation is that control respondents compared themselves with other people (both men and women), alternative interpretations are possible. Considering some of them, and referring to two studies we discussed above, they note (p. 197):

Recent literature finds some evidence that women and men’s SWB does react differently to comparisons with other people. This is true even when the outcomes over which these comparisons are made do not differ by gender (Ifcher et al., 2020). More generally, new evidence suggests that women and men differ in the aspects they consider when reporting SWB (Benjamin, Debnam Guzman, et al., 2023).

This quote, together with the authors’ findings, underline several recurring issues in using SWB data in experiments. First, as mentioned above and as demonstrated here by the main effect, the wording of SWB questions may dramatically matter for an experiment’s results. Second, the wording may matter *differently* for different demographic groups, who may, e.g., interpret the question or the response scale differently. That is, cross-group SWB differences may not necessarily reflect similar (indeed, any) differences in underlying well-being. Third, and closely related, experiments’ random assignment, combined with using a variety of SWB measures, could address some of these issues when analyzing the experimental manipulation’s main effect; however, comparisons across (non-randomly assigned) demographic

groups resurrects the familiar, thorny interpretation issues that plague observational studies. (Including the less-widely discussed possibility of differential self-selection, by demographic group, into the experiment.)

Our fourth example is Diaz et al. (2023). It is a follow-up, by the original coauthors, to the above 2020 study. In this follow-up paper they further develop a methodology for estimating social preferences based on *stated* SWB, i.e., *would-be* SWB, predicted ex ante—rather than on SWB reported ex post. In their lab experiment, participants ($N \approx 100$) report their would-be satisfaction with each of 21 (or more) payment profiles in which their own payment is fixed and another (anonymous, randomly assigned) participant’s payment varies. Specifically, participants are asked, at the top of the screen:

How satisfied would you be?

The question is followed by profile lines, each reading: “If you get \$20, and the other person gets \$[X]” (with [X] increasing by \$1 in each line). Each such profile line ends with a scale from 1 to 7, along with seven labels from “Extremely Dissatisfied” to “Extremely Satisfied.” Participants respond, for each profile, by (pseudo-)continuously moving a slider.

Their results generally support inequity aversion (larger for disadvantageous than for advantageous inequity). Participants also play economic games; inequity-aversion distributions estimated from choices in the games are different from those estimated from the “stated satisfaction” list. On which the authors comment (p. 12):

Differences between stated satisfaction and choices can be explained by considering SWB as an argument in the utility function (Benjamin, Heffetz, Kimball, & Szembrot, 2014). This would imply that stated satisfaction is an input into the decision-making process, and ultimate choices take other inputs into consideration, like context, monetary restrictions, and social norms.

Finally, the authors openly discuss potential demand effects (p. 16):

Also, we should not overstate the minimization of the role of reciprocity in our experiment, since participants may attempt to report higher satisfaction with equal payments because of assumed social norms of fairness and experimenter demand effects; experimenter demand effects would be less of a problem if participants each reported their satisfaction with only one payment-profile, though this would eliminate the ability to control for individual-satisfaction fixed-effects. That said, it could be argued that experimenter demand effects would bias stated-satisfaction patterns toward resembling efficiency concerns, as there is no tradeoff between own and other’s payment in our experiment.

The two issues discussed in these two quotes—the relationship between stated SWB and (stated or incentivized) choices; and the question of demand effects on either—have been increasingly studied by economists in the past 15 years or so (see, e.g., Benjamin et al., 2012, tellingly titled “What Do You Think Would Make You Happier? What Do You Think You Would Choose?”). It would be nice to see more empirical evidence such as Diaz et al. (2023) collected on both questions.

Our final lab experiment example was summarized in the introduction. Motivated by the question “Does prosocial behavior promote happiness?” Falk and Graeber (2020) extend previous research by considering both short- and long(er)-term effects. Their motivation (“Significance,” p. 6463) may suggest a hedonistic view of well-being:

Governments around the world increasingly acknowledge the role of happiness as a societal objective and implement policies that target national wellbeing levels. Knowledge about the determinants of happiness, however, is still limited. A longstanding candidate is prosocial behavior.

Each lab-session participant ($N \approx 300$) reports their happiness, mood, and self-image. They then choose between Lottery A: saving a life (of a person with tuberculosis) by triggering a targeted €350 donation with 60% probability and receiving €100 with 40% probability, and Lottery B: receiving €100 with 60% probability and saving a life with 40%. The lottery of choice is then drawn; the participant is informed of the result; and they report their happiness, mood, and self-image again. Reminder emails on the content and outcome of the lab session are sent one week and again three weeks after the session; in between, payment is transferred two weeks after the session. Finally, a follow-up (online) survey is conducted four weeks after the session, in which happiness, mood, and self-image are reported a third time.

Regarding their choice of measures, the authors write in their Supplementary Information Appendix (p. 2; emphasis, footnote 10, and formatting were added by us):

We chose our main measure of happiness to fulfill two requirements: It should be suited to capture both short-term as well as long-term variation in happiness, and it should be widely used and validated by the previous literature. Our measure is based on the Subjective Happiness Scale (Lyubomirsky & Lepper, 1999), also referred to as the General Happiness Scale. In particular, we use the first item, which is an assessment of the statement

In general, I consider myself:

with possible responses ranging from 1 = “not a very happy person” to 7 = “a

very happy person” on a 7-point Likert scale.¹⁰ Our measure of self-image is an assessment of the statement “I am a good person” on a 10-point Likert scale ... mood was elicited using the question

What is your mood at the moment?

and an 11-point response scale from 0 = “very bad” to 10 = “very good.”

The experimental effects they find on their happiness measure—the focus of their paper—were described in our introduction. (Their findings regarding mood and self-image effects differ and, in particular, decrease significantly from the session to four weeks later.) Here we add three observations. First, the happiness measure—“In general I consider myself (not) a very happy person”—could be interpreted as a question about personality (i.e., the type of person I consider myself) rather than about life evaluation or positive affect. In other words, as is often the case with SWB measures, it is not entirely clear what respondents understand the question to be asking and, hence, what exactly it measures. (For example, personality may be generally stable and hence less sensitive to the experimental manipulation.) Second, again as is often the case, demand effects are a concern here too; the authors’ attempt to reduce them through survey design (p. 6464): “The happiness, self-image, and mood questions formed part of a questionnaire including various other items so as to obfuscate the purpose of the study and to reduce experimenter demand effects.” While the experimental paradigm is such that the direction of potential demand effects is not as obvious here as in other experiments using SWB, the transparency of treatment allocation can increase the risk of such effects (two points we discuss in Section 5). Third, and related, the experimental paradigm may also reduce context effects (discussed in Section 4), as the only randomly assigned difference between participants is what outcome they get in a lottery they all participate in.

¹⁰The other three items on the Subjective Happiness Scale (Lyubomirsky & Lepper, 1999) are:

Compared to most of my peers, I consider myself:

Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?

Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you?

with response scales from 1 = “less happy” to 7 = “more happy” (second item) and from 1 = “not at all” to 7 = “a great deal” (third and fourth items).

3.2 Evaluating Programs and Policy Interventions

We now discuss four examples of past field studies where SWB measures play a central role. As mentioned in previous sections, relative to lab studies, (a) these field studies’ potentially life-changing interventions justify asking also more evaluative, and even eudaimonic, questions; (b) their scale means that they can collect more, and more detailed, SWB measures; (c) they can measure SWB many months or even years after the intervention; and (d) their authors typically view SWB as a complement to other measures that they collect, rather than as an exclusive alternative. At the same time, as we discuss in the two subsequent sections, some of these very features of field studies may also exacerbate context and demand effects.

Our first example is Devoto et al. (2012), who focus on SWB already in their title: “Happiness on Tap: Piped Water Adoption in Urban Morocco.” They conduct a field experiment in Tangiers, on households ($N \approx 800$) who are not directly connected to the city water system; have access to proper drinking water within walking distance; and can apply to a “Social Home Connections” program (in French, “Branchements Sociaux Individuels,” or BSI), offered by the city, allowing low-income households to purchase a direct piped connection on credit. Their treatment consists of a door-to-door information and marketing campaign about the BSI program, along with assistance in completing the application. It proves a success: six months after treatment initiation, 69 percent of treated households take on a loan for making the connection investment, compared with only 10 percent in the control group.

The authors’ motivation for including SWB measures among their collected outcomes is clearly stated early in their introduction. They are directly interested in well-being but, like other researchers, do not seem to commit to a particular view of what well-being is (p. 69):

The prime objective of connecting poor households to the drinking water network is usually to improve their physical health. Yet, over and beyond its direct effect on physical health, improved water access could have important effects on household well-being. By reducing the time burden of water collection, it not only frees up time that could be spent on additional leisure or production, it also removes an important source of stress and tension. On the other hand, if women face restricted mobility besides their daily trips to the public water source, a home water connection could cut off an important source of opportunities to socialize, and possibly reduce their well-being.

To assure data quality, the authors apply the de facto gold standard for avoiding demand effects (which they discuss as desirability bias): Their data collection is completely separated

from their intervention. In particular, the team that conducts the surveys is separate from the team that implements the information-and-assistance intervention and does not know about the research team’s relationship with the BSI program. (Participants are only informed about the full study design at the end of the study, when they are also asked for informed consent to let the researchers use their data.) The authors also try to minimize some potential context effects (which they discuss as priming and order effects) on SWB. For example (footnote 10, p. 91): “The questions about life satisfaction and mental well-being were not asked immediately following the questions about satisfaction with regards to water and house cleanliness, but in a much later part of the survey.”

Summarizing their findings, they write in their conclusion (p. 94):

After six months, this investment turned out to have significant private returns. While it freed up a lot of time for household members, it did not generate increases in labor market participation and income, but it led to important improvements in quality of life and welfare. In particular, it increased reported happiness and social integration. While the private returns are high, the public health impacts seem low, if not nonexistent. In the study context, household connections do not reduce the incidence of waterborne diseases compared to the existing system of public taps.

This “increased reported happiness” conclusion is largely based on four columns (4–7) in their “Table 11—Satisfaction and Well-Being.” Columns 4 and 5, respectively, show clear positive effects on average responses to two questions—the following question and a similar one about family’s situation (the survey is conducted in Arabic):

Are you under the impression that your (personal) situation improved since last year, at the same date?

(1 = “Yes, a lot,” 2 = “Yes, a bit,” 3 = “No, things haven’t really changed,” 4 = “No, the situation is worse.”) We note that the question asks about one’s (family’s) *situation*; participants are never asked about “happiness” per se and, as is often the case in the literature, the authors use this term rather loosely. Column 6 shows a statistically weak, positive effect on the share of respondents responding 5 and above to the ladder-question variant:

How would you classify your satisfaction with your life right now? If the top of the scale represents very high satisfaction and the bottom of the scale represents very low satisfaction, where would you place yourself?

(With a paper scale shown to the respondent: “1...2...3...4...5...6...7...8...9...10.”) Column 7 shows a statistically insignificant, positive effect on a “mental well-being” index: an average of (recoded) responses to three of the following eight questions (we add formatting, and underline the three questions included in the index):

During the last seven days, have you felt: Sad / Irritated / You got angry / Relaxed / Satisfied / Cheerful / Concerned/worried / Tired.

(1 = “all the time,” 2 = “often,” 3 = “sometimes,” 4 = “never.”) We note that the authors do not motivate their focus on these three emotions (sad, satisfied, concerned/worried), and do not report analysis of—or having collected—the other five.

Overall, the paper demonstrates the added-value potential of SWB outcomes as part of a high-quality data-collection effort in the field. In addition to the careful study design overall, the paper also demonstrates the particular importance of collecting different SWB measures, phrased differently and potentially measuring different aspects of well-being. It also provides examples of SWB questions that are explicit about the time frame they ask about (“since last year, at the same date,” “During the last seven days”)—in line with the recommendations in Benjamin, Debnam Guzman, et al. (2023). While the variety and detail of the collected data seem underutilized by the paper—especially for a paper on “happiness”—once SWB data are carefully collected and shared, future research can always return to them for further exploration.

Our second example is Ludwig et al. (2012). As their title clearly indicates, they aim to study “[n]eighborhood effects on the long-term well-being of low-income adults.” Their motivation is also clearly stated (without committing to a specific notion of well-being; p. 1505): “Nearly 9 million Americans live in extreme-poverty neighborhoods, places that also tend to be racially segregated and dangerous. Yet, the effects on the well-being of residents of moving out of such communities into less distressed areas remain uncertain.” (See also later in their paper, e.g., p. 1507: “[S]WB... was added to the long-term survey to be one of the key summary measures of the net impacts on families.”) They analyze SWB data collected as part of Moving To Opportunity (MTO)—a large-scale social experiment ($N \approx 4,600$ households) that helped randomly selected families relocate to less-distressed neighborhoods (using housing vouchers).

The authors report results from a long-term survey ($N \approx 3,300$) conducted 10–15 years after randomization. They estimate that an offer to relocate from a high-poverty to a lower-poverty neighborhood does not have statistically significant effects on measures of economic self-sufficiency (the point estimate is *negative*), physical health (positive) and mental health

(positive), and have a significant positive effect on the GSS happiness question:¹¹

Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

We make three observations. First, to the best of our knowledge, this pioneering study is a rare example of estimating causal effects on SWB a decade or more after a large policy intervention. While this long horizon may also introduce an additional SWB-related concern, namely, that participants’ use of the SWB response scale may adapt over time—and, moreover, may adapt *differently* across treatment groups—such adaptation would likely *attenuate* any estimated effects.¹²

Second, and closely related, the estimated effect on the GSS happiness question is large. To quantify, or “price” it, the authors follow standard practices in the large non-experimental literature on SWB in economics, and estimate so-called “happiness regressions”: SWB is the dependent variable and household characteristics are regressors. By comparing the coefficient on tract poverty rate with those on income and on tract share minority, the authors estimate that the effects of relocating from a high- to a lower-poverty neighborhood are larger than those of a fifty-percent increase in annual income or of relocating to lower-racial-segregation neighborhoods. Benjamin et al. (2024) show that in general, the interpretation of such estimates depends strongly on a list of assumptions regarding what exact notion of well-being the SWB question measures (we summarized some of them in Section 2). One important assumption relates to what time horizon participants have in mind when they report their SWB (e.g., how they interpret “these days” in the GSS question). However, because Ludwig et al. (2012) study the supposedly stable, long-term effects of an intervention, then, assuming the relationship between SWB and their regressors (e.g., income, neighborhood poverty rate) is also stable, interpretation may be more straightforward in this case.¹³

¹¹In selecting the GSS happiness question as their main SWB measure, they follow advice from Dolan and Metcalfe (2012), “Measuring Subjective Wellbeing: Recommendations on Measures for use by National Governments,” based on which Ludwig et al. (2012) conclude: “This type of happiness question yields results similar to those from questions about general life satisfaction; both provide global retrospective assessments of how people think their lives are going and are increasingly used to assess public policy impacts” (pp. 1505–1506). In addition, the authors’ measures of physical and mental health include some components that could also be viewed as SWB components. In particular, they report in their Supplementary Material that their physical health index includes self-reported health (fair vs. poor), and their index of adult mental health “consists of a psychological distress index score for the past month, lifetime depression, lifetime Generalized Anxiety Disorder, calm and peaceful during the past month, and normal sleep last night.”

¹²As Ludwig et al. (2012) report in their Supplementary Materials, they take some measures to address some usual concerns, such as potential demand effects. For example, interviewers are blinded to treatment group. At the same time, the long horizon makes attrition a concern. In particular, as the authors state in the first paragraph of their Methods section, to estimate effects they have to assume “that there is no selective attrition in measuring outcomes across groups” (p. 1507). They discuss effective response rate by treatment in their Supplementary Materials.

¹³Benjamin et al. (2024) analyze in detail the common example of pricing the SWB cost of unemployment

Finally, and also related, we again observe that in addition to researchers not being explicit about what notion of well-being they are interested in, they also use the relevant terms rather loosely. Devoto et al. (2012), for example, use the word “happiness” prominently throughout their paper, but they never mention it to participants in their survey questions. (As mentioned above, their main effects are found using questions on whether respondents’ *situation* improved; they do ask about, e.g., having been cheerful, and relaxed, in the last seven days—but they do not report any analysis of the answers.) Ludwig et al. (2012), on the other hand, refer in their paper to “well-being,” which they measure using questions about happiness, physical, and mental health. As long discussed in the literature, and as empirically demonstrated by some of the studies above, the actual wording of the SWB questions may matter. Collecting (and reporting on) more measures will improve future research—as will more clearly defining the terms used.

Our third example of a field study using SWB is Haushofer and Shapiro (2016). We summarized it in the introduction. It is more recent; it cites and builds on previous research; and, in several ways, it is another example of careful data collection and analysis—while still not committing to a specific view of well-being. The authors use an RCT to study the short-term response of poor households ($N \approx 1,500$) in rural Kenya to unconditional cash transfers (UCTs), with randomization of the treatment at both the village and household levels. Recipient gender (wife versus husband), transfer timing (a lump-sum transfer versus monthly installments), and transfer magnitude (US\$404 PPP versus \$1,525) are also randomized within the treatment group. Their motivation for studying SWB along more standard measures (p. 1976):

This setup allows us to assess the impact of UCTs and address a number of additional questions in the economics literature. First, we document the economic effects of UCTs on consumption (including temptation goods), asset holdings, and income, as well as broader welfare effects on health, food security, education, and female empowerment. We study in detail the effects on psychological well-being, including the stress hormone cortisol.

Similarly to the Morocco study above, “the survey team was kept distinct from the intervention team and denied any association when asked (although it remains possible that at least some respondents nevertheless suspected a connection).” The authors use a variety

in units of annual income (see their introduction and section 4.1). They show that if unemployment, as well as its assumed effect on SWB, are both temporary, while annual income and its SWB effects are more stable, then researchers need to know the time horizon captured by the SWB question. Otherwise, researchers do not know whether the estimated “price” of unemployment is a one-time cost to be paid once, or a recurring cost to be paid many times!

of SWB measures, purposefully elicited in a private interview “without the interference of other household members, especially the spouse.” These include the standardized total of the scores from: the 20 elements of the Center for Epidemiologic Studies Depression (CESD) Scale (Radloff, 1977), 16 elements of a custom worries questionnaire, and four elements of Cohen’s stress scale (Cohen et al., 1983); and the standardized numerical response to the World Values Survey happiness and life satisfaction questions:

Taking all things together, would you say you are “very happy” (1), “quite happy” (2), “not very happy” (3), or “not at all happy” (4)?

All things considered, how satisfied are you with your life as a whole these days on a scale of 1 to 10?

(From 1 = “Very dissatisfied” to 10 = “Very satisfied.”) They also report a “psychological well-being index”: A standardized weighted average of their negatively coded scores on the first three measures above, responses to the happiness and life satisfaction questions, and negatively coded log cortisol level.

They find that transfers had a substantial impact on all SWB measures (including the overall index), with larger impact for larger transfers, no overall difference between lump-sum transfers and monthly installments, and no significant difference between male and female recipients. They do not find an average treatment effect on cortisol.¹⁴ They also find statistically and economically significant effects on consumption, food security, and medical and educational expenditures; a marginally significant effect on a measure of optimism; and no effects on trust, locus of control, and self-esteem (they appear to consider the last four as additional “psychological well-being” measures). Interpreting these results, they write (pp. 2012–2013; the last sentence was quoted in our introduction):

In comparison to self-reported measures, our cortisol results suggest that either cortisol is noisier and more difficult to affect with interventions than self-reported measures, or that the self-reports may be subject to experimental demand effects. The fact that we observe significant positive effects on some of these variables but not others argues against the demand effect explanation; if people were telling the surveyors “what they wanted to hear,” we would have expected positive effects across the board. A further reason to think that noise plays a role is the fact that the treatment effects on self-report measures do not always have the same sign as those on cortisol; for instance, monthly transfer recipients have lower levels

¹⁴They do find that cortisol levels are significantly lower when transfers are made to the wife rather than husband, for lump-sum rather than monthly transfers, and for large rather than small transfers.

of depression than do lump-sum recipients, but higher levels of cortisol. Such discrepancies across variables are not unique to cortisol, but they illustrate a difficulty in using it as an outcome variable. In the absence of a gold standard measure that assesses well-being directly, it is not clear which variable should be given priority.

While not airtight, we would like to see similarly thoughtful discussions in all papers that use SWB along with other measures.¹⁵ Thoughtful discussions, in turn, are made easier by thoughtful data collection. Paraphrasing the last sentence in the quote above: In the absence of a direct measure of well-being, we would like to see researchers collecting a variety of outcomes, including a variety of SWB measures.

Our final example is the even more recent study by Allcott et al. (2020). The authors conduct an RCT ($N \approx 1,600$) to examine the impact of deactivating Facebook for the four weeks preceding the 2018 US midterm elections.¹⁶ They examine various outcomes, organized in nine outcome families and collected in a variety of ways. One family consists of seven SWB outcomes measured twice, in surveys at baseline and endline (happiness, life satisfaction, loneliness, depression, anxiety, absorption, and boredom); three outcomes measured in rotation via daily text messages (happiness, emotion, and loneliness); and an overall SWB index.¹⁷ The seven SWB survey outcomes are based on a variety of standard SWB measures, but: (a) they are rephrased to refer to the deactivation’s four weeks; and (b) in some cases they are subsets of standard multi-question sets, selected to focus on SWB aspects that may be most relevant to Facebook use. The authors’ stated interest in SWB seems to combine a utility-centered view typical of traditional economics—“We study the impact on individual utility via measures of subjective well-being” (p. 630)—with a direct interest in (experienced) well-being and mental health, traditionally more associated with psychology but, as demonstrated above, also shared by many economists, at least implicitly (p. 653):

SWB . . . outcomes are of interest because . . . many studies show cross-sectional or time-series correlations between social media use and well-being, and on this

¹⁵Like others, the authors never discuss context effects, to which their demand-effects argument—“if people were telling the surveyors ‘what they wanted to hear,’ we would have expected positive effects across the board”—may not easily apply. For example, if people were instead looking for contextual cues regarding what surveyors *meant* when asking an SWB question, we may not expect similar effects across the board, because different SWB measures could be subject to different resulting context effects. We return to this point in Section 4.

¹⁶Their impact-evaluation sample includes only participants who indicate, using BDM, a willingness to accept less than \$102 for deactivating Facebook for the four weeks. Treatment participants are offered \$102 for the deactivation, and control participants are offered \$0.

¹⁷Other outcome families measure news knowledge and political polarization, online and offline activity, post-experiment Facebook use, and much more.

basis researchers have speculated that social media may have serious adverse effects on mental health.

Their happiness survey measure is again based on the Subjective Happiness Scale (Lyubomirsky and Lepper, 1999; see our quote above from Falk and Graeber, 2020, along with our footnote 10, for the original four items). However, they average the responses of the first *two* items, and they adjust them for the four-week time frame. We note that this adjustment also replaces the “In general, I consider myself” language, suggestive of personality, with language potentially more suggestive of emotion/evaluation:

Over the last 4 weeks, I think I was:

1 = “not a very happy person ... 7 = “a very happy person”)

Over the last 4 weeks, compared to most of my peers, I think I was:

(1 = “less happy” ... 7 = “more happy.”)

Life satisfaction is measured as the sum response to an adjusted version of the three satisfaction questions from the Satisfaction with Life Scale (Diener et al., 1985):

Below are three statements that you may agree or disagree with. Indicate your agreement with each item and please be open and honest in your responding.

In most ways my life during the past 4 weeks was close to ideal.

The conditions of my life during the past 4 weeks were excellent.

During the past 4 weeks, I was satisfied with my life.

(“Strongly disagree,” “Disagree,” “Slightly disagree,” “Neither agree nor disagree,” “Slightly agree,” “Agree,” “Strongly agree.”) For the loneliness survey measure, they use their adjusted version of the Three-Item Loneliness Scale from Hughes et al. (2004).¹⁸ Their last four survey measures—depressed, anxious, absorbed, and bored—are questions from the European Social Survey well-being module (Huppert et al., 2009) that they again adjust for their context:

Below are some ways you might have felt or behaved in the past 4 weeks. Please tell us how much of the time during the past 4 weeks:

¹⁸The Three-Item Loneliness Scale is a sum of three questions, which the authors adjust to:

How often did you feel [that you lacked companionship]/[left out]/[isolated from others] over the past four weeks?

(1 = “Hardly ever,” 2 = “Some of the time,” 3 = “Often.” The original three questions in Hughes et al. (2004) are: “How often do you feel [that you lack companionship]/[left out]/[isolated from others]?” with the same response options.)

you felt depressed
you felt anxious
you were absorbed in doing something worthwhile
you felt bored.

(1 = “None or almost none of the time” ... 4 = “All or almost all of the time.”) Their three rotating text-message questions (on happiness, emotions, and loneliness):

Overall, how happy do you feel right now on a scale from 1 (not at all happy) to 10 (completely happy)?

What best describes how you felt over the last ten minutes?

How lonely are you feeling right now on a scale from 1 (not at all lonely) to 10 (very lonely)?

(The second question above offers a list of emotions to choose from; responses are coded 1 if a positive emotion is chosen and 0 otherwise.) Finally, an SWB index weights each measure by the inverse of the covariance between variables at endline (for survey variables) or over the survey period (for text-message variables), as described in Anderson (2008).

The authors find that deactivation mostly caused small improvements in endline SWB, insignificant improvements in text-message SWB (also not significant in any of the four weeks separately), and a small improvement in the overall SWB index. Describing the effects as “small” is based on comparisons with three benchmarks that the authors offer and thoughtfully discuss: the original response units; a meta-analysis of randomized evaluations of psychology interventions; and, as above, coefficient ratios from “happiness regressions.” We commented above on the difficulties in interpreting such coefficient ratios (and provided a cautionary example in footnote 13). Interpretation may be easier here because the authors’ SWB questions clearly state to participants what time horizon they are asked about; but it may also be more complicated because, as the authors observe (p. 633): “A longer period without Facebook might have ... either more or less impact on subjective well-being.”

The authors devote a subsection (IV.F) to demand-effect concerns, observing (p. 666): “Most of our outcomes are self-reported, and it would have been difficult to further conceal the intent of the randomized experiment.” As we discuss in Section 5, they collect and provide suggestive evidence that their findings are not mostly driven by demand effects (by directly asking participants about the researchers’ agenda). They also compare their estimated effects on SWB with correlational studies (their effects are smaller), and with participants’ descriptions in follow-up interviews and free-response questions (which suggest substantial heterogeneity).

We would like to see such comprehensive discussion of SWB findings in all future work. Which in turn requires comprehensive data collection. Indeed, this study demonstrates much that we recommend in this regard: (a) It collects a long list of measures, (b) asking about a variety of SWB types and aspects, (c) during several different, clearly stated time horizons, and (d) collected at different frequencies (and using different media).¹⁹

In addition, the authors’ *findings* demonstrate the practical relevance of these recommendations. For example, as mentioned above, their text-message SWB results are never statistically significant—unlike some of their survey SWB results. Remarkably, both include questions that ask specifically about being happy: a “happy person” and “more happy” than peers, in the survey; feeling “happy ... right now” in text-messages. Again, such findings are a reminder that details like specific wording (emphasizing, e.g., emotion, evaluation, time horizon, domain) and context (timing, location, frequency, medium) may all matter.

Finally, beyond examining a variety of SWB measures, the authors also return to compare them as a group with more traditional economic outcomes. For example, echoing the discussion in Haushofer and Shapiro (2016) regarding the discrepancy between SWB and cortisol, Allcott et al. (2020) reach somewhat analogous conclusions regarding an apparent discrepancy between SWB and BDM valuations (p. 633):

It is not entirely clear whether one should prioritize the [SWB] survey measures or monetary valuations as normative measures of consumer welfare. Benjamin et al. (2012) suggests that [SWB] measures like ours are not a complete measure of what people are trying to maximize when they make decisions, but ... other studies make clear that monetary valuations are not closely held and can be easily manipulated. We think of these tensions as fodder for future research.

¹⁹In a follow-up paper investigating digital addiction, Allcott et al. (2022) demonstrate a significantly simplified SWB-data-collection approach. (Their intervention provides either monetary incentives or a special-purpose app to help participants reduce time spent on popular phone apps.) They conduct repeated surveys, at three-week intervals. In each, they measure SWB as the (unweighted) sum of seven 1–7 scores (after reverse-coding anxious, depressed, and easily distracted) obtained from the following question (reformatted by us):

Please tell us the extent to which you agree or disagree with each of the following statements. Over the last three weeks ... I was a happy person / I was satisfied with my life / I felt anxious / I felt depressed / I could concentrate on what I was doing / I was easily distracted / I slept well

(“Strongly disagree” ... “Neutral” ... “Strongly agree.” They also measure addiction via text-message questions three times a week; some addiction questions, e.g., about feeling down or losing sleep, could be considered SWB questions—as could another question they ask, about smartphone use having made life better or worse.)

4 Context Effects on SWB

Revisiting our cautionary tale from the introduction, imagine that you are out to find out, experimentally, how to meaningfully improve people’s overall well-being. You run a bunch of RCTs, each randomly varying something that you hypothesize may be big enough to matter—access to healthcare, wealth, social freedoms, etc. For outcome variable, you ask respondents where they stand on an 11-point scale, where 0 is “the worst possible life for you” and 10 is “the best possible life for you.” As a sanity check, to make sure your SWB measure is sensible, imagine that you also run an RCT where you vary something that you are pretty sure should *not* affect ladder responses: prior to the SWB question, you ask a randomly chosen half of your respondents two short political-opinion questions (e.g.: Do you approve of the way the President is handling his job? Are you satisfied with the way things are going in the Country?).

Imagine that to your surprise, you find, in a large US sample, that asking the political questions actually causes the average ladder response to drop by an amount comparable to that caused by the 2008 Global Financial Crisis, or to that expected by making all respondents unemployed. What will you have learned about the determinants of Americans’ well-being?

4.1 The Gallup Evidence

Deaton (2012) and Deaton and Stone (2016) report the worrying findings in the previous paragraph, as well as additional ones, from what is essentially the political-questions RCT described above. It was conducted by the Gallup Organization after noticing an unintended large context effect of two political questions on ladder responses in its Gallup-Healthways Well-being Index (GHWBI) poll. The RCT was conducted on a very large sample (around 1,000 respondents *a day*—500 in each treatment group) for 75 days in 2009.²⁰

While Deaton (2012) conjectured that this political-questions context effect moved SWB responses by negatively affecting respondents’ mood, Deaton and Stone (2016) conduct additional analysis, report additional findings, and suggest more nuanced explanations. One notable finding is that context effects can in some cases change SWB rankings across groups: Black respondents, typically the group with the lowest ladder responses, report higher SWB than either Whites or Hispanics when the SWB question is preceded by a question about the performance of President Obama. (Other group rankings, based on age, gender, or income,

²⁰The political questions were “Do you approve or disapprove of the way that Barack Obama is handling his job as president?” followed by “In general, are you satisfied or dissatisfied with the way things are going in the USA?” The ladder question immediately followed in one condition, and was the survey’s first question in the other condition.

are not found to be affected.) One of their conclusions is that the lower ladder scores “come from both thinking the country is going in the wrong direction and being asked to report the fact.” Overall, they interpret their findings as “consistent with the argument that, when asked a difficult question to which they have no ready answer, [respondents] 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’.”

As Deaton and Stone (2016) note, context effects have a long history in psychology. In a comment on Deaton and Stone (in the same journal issue), Lucas et al. (2016) aim to put the above findings in this broader context. First, they readily agree, as many others before them have, that responding to SWB questions is not straightforward: “People may not be able to accurately assess their internal feelings, they may wish to hide or distort these feelings for an external audience, or they may have trouble translating their internal state to a discrete response option.”²¹ Second, they note that “Deaton and Stone’s study is notable because it is one of just a few examples of practically important context effects found in reasonably large samples.” They therefore ask why the effect emerged in these data (“given the weak context effects that are typically reported in the literature”). Understanding the mechanism behind the effect could suggest ways to mitigate it.

Third, Lucas et al. (2016) attempt at an answer. They argue against two potential mechanisms, salience and mood. On salience, they note that Deaton (2012) shows that a single buffer question between the political questions and the ladder question (“Now thinking about your personal life, are you satisfied with your personal life today?”) was effective in mitigating much of the context effect; this, they argue, rules out salience because “that buffer question would not reduce the salience of that information or provide a long enough period of time for the salient information to fade from memory.” On mood, in addition to an argument similar to that used to rule out salience (i.e., the above single question is unlikely to reduce a mood effect), they also note that past research establishes that substantial mood changes are rather difficult to induce, even through targeted manipulations.²² Lucas et al.

²¹Lucas (2018) further explains (p. 47):

People may misremember their affective experiences, may disproportionately weight certain irrelevant sources of information when deriving a cognitive judgment of life satisfaction, may be unable to translate an internal feeling to a meaningful response on an experimenter-provided scale, or may be unwilling to provide an honest response when questioned about their well-being.

²²In the words of Lucas et al. (2016):

Psychologists have used mood inductions for decades, and even relatively powerful inductions (such as watching extremely sad movies or writing about the most important negative event in one’s life) have only moderate effects on current mood. Given what we know about the

(2016) therefore raise a third possibility (p. 874):

[T]he inclusion of the political questions actually changes respondents' interpretation of the global well-being question itself. As survey researchers have known for years (see, e.g., Schwarz, 1999), respondents do not simply respond to the literal meaning of the questions that are asked of them. Instead, they use contextual features of the survey or setting to infer what it is that the researchers really mean. In many (perhaps most) cases, these inferences likely match with the goals of the researcher; but in some cases they may not, at least for some respondents.

They further illustrate (p. 874):

Specifically, it may be reasonable for respondents to assume that researchers who first ask about the opinions of President Obama and then ask about direction of the country (especially researchers from the Gallup Organization, a known political pollster) are really interested in the way that the political situation influences the respondent's life as a whole. In other words, respondents, or at least some respondents, may have interpreted the Cantril ladder as a third question about their global evaluation of the way things are going in the country.

They also note that this interpretation may be consistent with the effect of the single buffer question, which would, presumably, shift respondents' focus to their personal life.

If the large context effect in the Gallup data is driven by respondents' use of the political questions as cues suggesting what the researchers really mean by their SWB question, then this context effect may be closely related to demand effects—a point we return to in the next section.

4.2 Beyond Gallup

The above discussion of the Gallup evidence, as well as more recent discussions in the psychology literature of the broader body of evidence, could be summarized as follows. While acknowledging the important, large-sample Gallup evidence of a context effect, psychologists still seem to generally view it as a relatively rare exception rather than as one representative

types of things that lead to substantial changes in momentary mood, it is not plausible that these two questions could change mood so much as to then influence well-being judgements in a substantial way (and it is important to remember that the effect of political questions on life satisfaction would be a multiplicative effect of the questions on mood and then mood on the judgement).

example out of many. Their skepticism may be understandable: early findings suggesting large context effects on SWB, having had large impact within psychology in the 1980s and 1990s, subsequently failed to replicate upon closer scrutiny—as recounted, for example, in Handbook of Well Being chapters by Lucas (2018), Robinson and Klein (2018) (in a section tellingly titled “Are Life Satisfaction Judgments Plagued by Context Effects?”), and Pavot (2018), among others. Perhaps most notably, Schimmack and Oishi (2005) failed to find large question-order effects on reported life satisfaction in several replication studies, and Yap et al. (2017) failed to find large mood effects in nine replication studies with substantially larger samples than the original studies. It appears that all agree, however, that more research is needed.

Such research is beginning to appear. Almost a decade after the Gallup political-questions experiment, under very different political circumstances in the US (January 2018), Stone et al. (2022) return to investigate the original context effect and ways to mitigate it. They conduct two experimental studies, one on workers on Amazon’s Mechanical Turk (MTurk) and the other on respondents in the Understanding America Study (UAS). Both studies replicate the original finding of a large context effect when the ladder question is preceded by a political question (this time, it is a single question—“In general, are you satisfied or dissatisfied with the way things are going in the USA?”—with binary response options—“Satisfied” and “Dissatisfied”).

Consistent with the long quotes above from Lucas et al. (2016), the first study in Stone et al. (2022) also finds that among several manipulations investigated, the most effective in mitigating the political-question effect is, indeed, a simple explanation providing information on what the ladder question means.²³ To be effective, the explanation has to come immediately before the ladder question; a manipulation where it comes before the context-inducing political question is less effective. Their second study ended up with much smaller samples, and failed to replicate the large mitigating effect of the explanation manipulation.

Encouraged by the mitigating effect of the explanation manipulation in their MTurk study, Stone et al. (2022) suggest that future work could try to further improve the explanation, perhaps by making it longer or more detailed. It is not clear why the mitigating effect

²³ The explanation reads as follows:

In just a moment we would like you to tell us about how you feel about your current life. From prior discussions with people, we know that many different aspects of life are considered in this rating. People often reflect on their physical and mental health, on their financial situation and standard of living, on their children and family, on their friends, on their employment/retirement, the amount of leisure time they have, and on where they live. You should think about these and other aspects of your life in answering the following question. Check the answer that gets closest.

failed to replicate in the UAS study. The authors mention potential explanations for the mixed results—the different samples, with different demographics, different attentiveness, different formats of the context-inducing question—and recommend additional future investigation of the explanation manipulation. If, as they speculate (in line with the apparent evidence from psychology), “variables tapping relatively vague or imprecise information that demand considerable reflection and judgment may be particularly susceptible to context effects,” then a manipulation aimed at making the information sought more precise and at reducing the demanded reflection and judgement is a natural direction to investigate. We examine this idea next.

4.3 Implications and Tentative Advice

An experimental treatment is always and everywhere also a context(-inducing) item. No matter what an experimental condition manipulates—resources, payouts, prices, endowments, information, expectations, incentives, the strategic environment, others’ behavior, or any other aspect of the subject’s (actual or perceived) situation—the experimental condition also manipulates the context under which outcomes are measured. If the large context effects reviewed above arise either because a context item “shifts the attention or focus of the respondent in a way that impacts the dependent variable”—the explanation emphasized by Stone et al. (2022), referring back to ideas by Tourangeau (1992)—or because context items are used by respondents to (mis)“infer what it is that the researchers really mean”—the explanation emphasized by Lucas et al. (2016), referring back to Schwarz and Strack (1999)—then every experimental manipulation, by also being a context item *by definition*, could potentially drive context effects on SWB measures.

However, all context(-inducing) items are not the same, and their potential unintended effects could vary considerably. On the one hand, a field treatment in which researchers provide participants with easier access to better housing, or block their access to social media may, relative to a control, both shift attention and focus towards these changes and (independently) (mis)suggest to respondents what it is that those same researchers really mean when they ask a general SWB question.²⁴ Since treated subjects are likely keenly aware of how much time, effort, and other resources the researchers invested in carrying out the treatment, it would be surprising if upon responding to an SWB question such context details would neither shift respondents’ focus and attention nor be used by respondents to

²⁴These context effects are separate from demand effects, covered in the next section, where respondents may use the context to try to infer what it is that the researchers *would like them to respond*. In particular, as in the Gallup example and in contrast with demand effects, respondents’ perceptions of the researchers’ agenda or hypotheses may not be a relevant driver of context effects. Indeed, Gallup is known as a *nonpartisan* political pollster.

try to infer what the researchers are “really” asking about.

Similarly, though perhaps to a lesser extent, a lab manipulation that alters the conditions under which participants make choices also alters context (by definition), and is hence a potential context item. It would be surprising, for example, if a treatment explicitly informing participants about *other* participants’ payments did not shift attention and focus in a way that could affect responses to an SWB question asked immediately after that information is realized, or if it were not used by respondents to try to infer what the SWB question is “really” asking.

On the other hand, when participants in all experimental groups are subject to exactly the same information, and the experimental condition is the realized outcome of a random variable—e.g., the result of a binary 60%-40% lottery (as in Falk and Graeber, 2020)—potential treatment-dependent context effects do not seem to be a big concern. (Of course, treatment-*independent* context effects could still be large, as the experimental setting itself may both shift attention and suggest to respondents what the researchers are “really” asking.)

What can be done? Fortunately, the very nature of the problem also suggests potential solutions. They require further research; we hope that future studies will at least experiment with them.

First, within a given experimental design, researchers should think about context effects when designing their SWB questions. As Lucas et al. (2016) put it (p. 875): “if the context effect emerges from these subtle influences on respondent interpretation, then it may be relatively easy for survey designers to communicate the true intention of the question.” This advice, which is consistent with the finding in the first study by Deaton and Stone, 2016 (see footnote 23 above for the full text of their communication), is also generally consistent with the advice, given in other contexts, to design clearer SWB questions. For example, as mentioned above, Benjamin, Debnam Guzman, et al. (2023)—who are concerned with issues unrelated to context effects—recommend authoring SWB questions that make explicit the time frame and social circles respondents are asked about. While some of the example studies discussed in Section 3 already do that, to our knowledge none of them attempts to also “communicate the true intention of the question,” as Lucas et al. (2016) recommend. We join this recommendation, as well as similar ones by Deaton and Stone (2016): We would like to see such attempts in future experiments, if only for the purpose of better understanding their effects.

Second, as discussed above, researchers should think about context effects when designing their experiments. The analysis by Deaton and Stone (2016) leaves open the question whether “easier” SWB questions (e.g., about emotions felt yesterday) are less affected by

context effects; their evidence is inconclusive.²⁵ If future research finds that such “easier” questions are indeed less susceptible to context effects, that would be good news for lab experiments, which use such more-focused SWB questions anyway (for other reasons discussed above). On the other hand, they note in their conclusion 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.” But unlike the routine large-scale surveys they have in mind, in experiments, which are our focus here, the context is often the very treatment. Starting a post-treatment survey with an SWB question eliminates any systematic context effects from preceding questions—but not from preceding treatments. To eliminate those, researchers have to go beyond SWB question design (e.g., question order, framing, and explanation); as discussed above, researchers have to carefully think about *treatment* design (e.g., à la Falk and Graeber, 2020) and *data-collection* strategy (e.g., à la Devoto et al., 2012 and others).

5 Potential Experimenter Demand Effects on SWB

Arguably the mother of all cautionary tales—indeed, a cautionary horror story—is the now-(in)famous experiment on obedience to authority by Milgram (1963). A participant (“teacher”) is instructed to administer increasingly severe electric shocks to another person (“learner”) whenever the “learner” answers questions incorrectly. The “teacher” believes the “learner” is a fellow participant. No monetary incentives are involved (beyond a show-up fee). We leave aside the ethical controversies, competing psychological interpretations, and many important, nuanced details of the original study and its many follow-up variations. For our purposes in this chapter, the study and its replications reveal the shocking extent to which individuals may follow experimental instructions. Disturbingly, many participants in these studies choose to follow instructions even when they are in clear conflict with the participants’ own moral beliefs. As a result, many participants engage in actions within

²⁵Deaton and Stone (2016) also find a large average effect of the political-questions treatment on a standard-of-living question. They conclude (p. 865):

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.

the experiment that they would not have dreamed of engaging in outside the experimental setting.

Around the same time, Orne (1962) introduces the role of demand characteristics in psychology experiments. Demand characteristics refer to cues or features of an experiment that convey information to participants about the purpose or hypothesis of the study, potentially influencing their responses and behavior. Of course, obeying an experiment’s explicit instructions is rather different from responding to implicit cues regarding an experiment’s purpose. But such cues could well be construed by participants as an implicit instruction—a suggestion, a nudge, an invitation—to behave in ways that the experimenter is believed to welcome. Findings à la Milgram and demand characteristics à la Orne may thus be closely related. In both, experimenters and experimental procedures affect participants’ behavior, including potentially influencing them to go against their ethical judgements regarding what, outside the experiment, they may consider right or wrong.

In this section we ask: To what extent might participants, aware of being studied, alter their SWB reports, either in response to potential cues about the experimenter’s aims or, perhaps more generally, in an attempt to be responsive (or helpful, polite, appreciative, or grateful) to the experimenter? Put simply: To what extent could experimenter demand effects, broadly defined, influence SWB responses in experiments?

5.1 Recent Non-SWB Evidence

In contrast with the much studied (and argued about) context effects on SWB reports (see previous section), we are not aware of existing research directly studying demand effects on SWB. What can we learn from recent investigations of demand effects on other outcomes?

As in the rest of this chapter, we focus on work in economics. Recent studies attempt to intentionally induce demand effects in the lab, and assess their effects on behavior. Their findings are mixed. Bischoff and Frank (2011), for example, study a “solidarity game,” where participants ($N \approx 180$) can choose to share money they randomly won with participants who did not win. In each of two experimental conditions, instructions are presented to participants using one of two prerecorded instructional films, in which the same actor was asked to induce, through their presentation, either a high or a low level of solidarity. According to the authors: “The actor was paid the customary fee. In addition, we promised him a case of high-quality apple juice (a regional delicacy) if we were to observe significant differences results [sic] from the two treatments. The actor used visual as well as vocal cues, but by no means overacted.” They find little evidence of different participant behavior across the two treatments.

In another example, Tsutsui and Zizzo (2014) develop an individual measure of susceptibility to demand effects, elicited at the end of an experiment on group status and trust. Their measure presents participants ($N \approx 300$) with a series of choices among lotteries, some of which are dominated; “the dominated option was characterized by a smiley face and a sentence stating that ‘it would be nice if some of you were to choose such an option’.” The authors consider participants who choose more of the dominated lotteries more prone to demand effects. However, they observe no correlation between this measure and participants’ behavior earlier in the experiment.

At the same time, the potential for meaningful demand effects remains real. Zizzo (2010) reviews older work studying potential demand effects in economic experiments. One of his examples is Brañas-Garza (2007), who uses framing to explore the role of helping behavior in dictator games. He finds that adding to the instructions, in bold characters and in a salient, centered position, the cue “REMEMBER that he is in your hands” (translated from the original Spanish) regarding the other participant, appears to increase dictator giving. While his samples are, especially by current standards, small ($N \approx 90$ across two treatments $\times$ two settings), the effect appears large, and more so in a classroom setting run by the professor (for extra credit points) than in a regular double-blind lab version (for money), raising the further possibility that obedience or anonymity could play a role.

More recently (and with a significantly larger sample), Cilliers et al. (2015) study dictator games in Sierra Leone ($N \approx 700$ participants from 60 villages: 30 control and 30 treatment), and randomly vary whether an experiment “supervisor” is a local man (control) or an American white man (treatment).²⁶ The authors find that the foreigner’s presence increases player contributions by around 19 percent (in their preferred specification), albeit statistically weakly. Based on this and additional evidence, they conclude (p. 413):

Our findings suggest that players act based on their perception of this individual as a white foreigner. We find results consistent with two types of demand effects. First, players appear to give more because they perceive this is what he wants them to do. For example, we find that real-world contributions strongly predict giving in games in the games [sic] where the white foreigner is absent, but do not in the games where he is present. As such, giving is less indicative of true generosity in the white foreigner’s presence.

Second, we find that players from more aid-exposed villages and households of village leaders give less in the presence of the white person. When asked about

²⁶Both “supervisors” are young, college-educated males in their 20s. “Their sole tasks were to take notes and distribute money to players in the beginning of each game. They did not talk during the administration of the game, or otherwise interact with game participants in any way.”

the purpose of the games, they are also more likely to report that the games where he is present test them for aid suitability. This is consistent with the idea that those knowledgeable about aid perceive the games as a form of means-testing and give less to signal they are poor and in need of further assistance.

This conclusion, while referring to a specific, imprecisely estimated main treatment effect on giving, echoes our more general worries regarding potential effects on SWB reporting—on which we do not have direct evidence. SWB data are often collected by, or on behalf of, foreign researchers, NGO workers, and others who may most naturally appear to participants as well-meaning individuals, interested in investigating the effectiveness of aid programs and other interventions. Moreover, participants may realize that these individuals, or the institutions that sent them, designed these interventions with the clear objective of improving these very participants’ lives, and invested heavily in carrying out these interventions. An SWB question from such individuals in the context of post-treatment program evaluation may therefore not be most naturally perceived by participants as a request for disinterested reporting. Rather, it may come across as asking: How did we do? Are we in the right direction? Should we keep trying?

Most recently (and with even larger samples), de Quidt et al. (2018) aim to more systematically study what they term “demand treatments.” They conduct seven online experiments (total $N \approx 19,000$ participants), in which they examine eleven canonical tasks that include, among others, the dictator and ultimatum games, and simple time, risk, and ambiguity preference-elicitation tasks. On three of the eleven tasks they also examine incentivized versus hypothetical choice. They employ two treatments, “weak” and “strong,” which they describe as follows (p. 3257):

“Weak” demand treatments signal an experimental hypothesis to our respondents: we tell them “We expect that participants who are shown these instructions will [work, invest, ...] more/less than they normally would.” We believe that these treatments are likely to be more informative than implicit signals about demand in typical studies, so in our view these bounds will be sufficient for most applications. Our “strong” demand treatments go further, telling participants “You will do us a favor if you [work, invest, ...] more/less than you normally would.”

They use both positive and negative demand treatments (conveyed to participants using “more” or “less”; see quotes above).²⁷ They find that the response to the weak demand

²⁷For example, the weak positive demand treatment in the dictator game is: “We expect that participants

treatments is modest (averaging around 0.13 standard deviations), and “interpret these results as suggesting that demand effects in typical experiments are likely to be small” (p. 3268).²⁸ However, for the strong demand treatments they observe significantly larger reactions (averaging 0.6 standard deviations), on which they further comment (p. 3268): “...thus researchers are right to pay close attention to potential demand effects in their studies.”²⁹

This (implied) recommendation to pay close attention to potential demand effects appears characteristic of current views among these and other experimentalists in economics, who at the same time express optimism. For example, in a recent review, de Quidt et al. (2019) conclude that while demand effects could potentially be significant, present best-practice experimental design—which has been developed partly to avoid them—appears to be doing its job. As evidence, they highlight the absence of known examples of demand effects that reversed qualitative (directional) effects in economic experiments. They also provide evidence for the adoption of best-practice design in recent years, in order to show that these norms are indeed generally accepted.

5.2 What Can the Evidence Tell Us Regarding SWB?

The above experiments studying potential demand effects differ from the example experiments measuring SWB reviewed in Section 3 in three important ways: (i) They largely rely on game-based experiments involving actions, rather than on self-reports of emotions and life evaluations; (ii) said actions are often incentivized (but see counterexample below); and (iii) the stakes are typically low (we are not aware of demand-treatment experiments within potentially life-changing field studies). We now discuss these three differences.

First, compared with actions in games, correctly inferring a researcher’s hypothesis may be easier with SWB. Consider a randomly assigned condition or outcome, e.g., winning money in the lab or receiving an unconditional cash transfer in the field. A participant may more easily infer that it is hypothesized to improve their SWB than that it is hypothesized to change their action in a risk-preference task in a particular direction.³⁰ Indeed, de Quidt

who are shown these instructions will give more to the other participant than they normally would”; the strong negative demand treatment is: “You will do us a favor if you give less to the other participant than you normally would.”

²⁸For a similar, concurrent study in political science, with limited evidence of even modest effects and therefore an even stronger conclusion, see Mummolo and Peterson (2019).

²⁹de Quidt et al. (2024) review recent work that uses the demand treatments approach of de Quidt et al. (2018) and conclude, similarly to the original paper, that “... across studies the effect of deliberately inducing experimenter demand is generally modest.”

³⁰Similarly, Mummolo and Peterson (2019) argue that in political-science survey experiments, participants’ ability to both correctly infer experimenter intent and act accordingly may be limited. In particular, like inferences from some economic hypotheses to specific actions in games, inferences from some political-

et al. (2018), who make it easy on game participants by explicitly informing them of the experimenter’s intended direction of action change, find consistent demand effects (though modest and mostly not statistically significant in their weak treatments).

Second, SWB outcomes are never incentivized—an issue that goes well beyond demand-effect concerns. Indeed, traditionally, economists have been notoriously skeptical of any non-incentivized outcomes, in any context. A full discussion is well outside the scope of this chapter; fortunately for economics as a science, this traditional taboo has long been lifted, and a wider set of evidence is now welcome by most economists. On the specific question of demand effects, de Quidt et al. (2018) test their demand treatments on both hypothetical and incentivized choices, and find similar effects (they look at a dictator game, a risk-aversion task, and a time-preferences task). However, they find a much larger effect of their strong demand treatments in an effort task when it is unpaid versus paid. They conclude (p. 3929): “The mixed evidence on responsiveness to incentives is somewhat surprising. One possibility is that our incentivized choices still involve relatively low stakes, and that we would see a difference at higher stakes.”

Finally, one may worry that to the extent that participants would like to help or please a researcher, their tendency to do so may depend on, e.g., how grateful they feel towards the researcher or how much impact they feel that the research has had on their lives. Consider these demand treatments from the studies above: “It would be nice if some of you were to [do X]” and “We expect that participants . . . [will do X].” Even if such treatments do not cause large demand effects among small-stakes lab participants, they could potentially cause large effects—perhaps especially on self-reported outcomes, and particularly on SWB—among recipients of large monetary transfers or access to housing in better neighborhoods. Indeed, a participant may feel forever indebted to an experimenter who offered them better housing, but not at all to one who provided them with a few dollars. This concern is less prominent in low-stakes online or lab experiments, which may contribute to an apparent optimistic outlook in the literature regarding online experiments.

These three differences in settings, between experiments that have been used to study SWB outcomes and those that have directly studied demand effects, may suggest that po-

science hypotheses to specific reported opinions and views may not be obvious.

We demonstrate the difficulty using one of their own experiments (Mummolo and Peterson, 2019, p. 525; they have several other experiments, including a somewhat similar one, with different results). Control participants are instructed: “You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read.” All treated participants are informed in addition: “The researchers conducting this survey expect that individuals are more likely to choose a news story if it is offered by a news outlet with a reputation of being friendly towards their preferred political party.” Some of the treated are even incentivized; they are told in addition: “If your responses support this theory, you will receive a \$0.25 bonus payment!” The authors find that these treatments do not affect behavior, including mere guesses of the experimenter’s hypothesis later in the survey.

tential demand effects may be an even more important consideration in studies involving SWB. Moreover, some similarities across the settings may suggest still further caution. For instance, one might wonder whether field experiments carried out in foreign developing countries, such as Devoto et al. (2012) and Haushofer and Shapiro (2016), may be more susceptible regarding their SWB outcomes to one or both types of demand effects discussed by Cilliers et al. (2015) and quoted above (see p. 38). We also note that while we did not discuss SWB effect sizes in the experiments in Section 3, these are often comparable, in standard deviations, to the effects of the weak demand treatments in de Quidt et al. (2018), and smaller than the effects of their strong demand treatments. Finally, and related, one may wonder whether some SWB questions may *themselves* be perceived as demand treatments, by at least some participants in some experiments. And if so, whether they are perceived as weaker or stronger than, e.g., the weak and strong demand treatments of de Quidt et al. (2018).

5.3 Practical advice

Recent work in economics that provides useful advice regarding demand effects includes the Handbook chapter by de Quidt et al. (2019), the paper by Haaland et al. (2023) on designing information-provision experiments, and the guide by Stantcheva (2023) on conducting survey experiments. Their recommendations are mostly based on two main ideas: (i) Make it difficult for participants to guess the experimental hypothesis and (ii) diminish the likelihood of experiment participants complying with such hypothesis (for example, due to obedience to authority figures).

Some of these recommendations, such as those regarding incentives, are not relevant for SWB. Of the others, we speculate that the most relevant are, first, maintaining participant anonymity. In the words of de Quidt et al. (2019): “Anonymity helps minimize the interaction between the experimenter and the participant and thus reduces temptation to deviate from the true preferred choice to please the experimenter.” Second, obfuscated follow-up studies, with the same pool of participants who are not aware that the studies are related, have long been recommended (for recent implementation and detailed discussion, see, e.g., Haaland and Roth, 2020 and Haaland and Roth, 2023); as mentioned in Section 3, Devoto et al. (2012) and others carefully implement an obfuscated follow-up survey to measure SWB and other outcomes. Third, obfuscated information treatments add information or tasks that are not the objective of the experiment and can therefore help mask the objective from participants. Ifcher et al. (2020), discussed above, incorporate elements from the Basic Psychological Need Satisfaction Scale (BPNSS) in the middle of their survey. More gener-

ally, hiding SWB questions among other questions may serve a similar purpose. We warn however that questions preceding an SWB question may result in strong context effects, as documented in Section 4. Fourth, researchers are advised to measure, at the end of an experiment, participants’ beliefs about its purpose. Allcott et al. (2020), discussed above, ask participants “Do you think the researchers in this study had an agenda?” and use the responses to look for (and not find any) suggestive evidence of potential demand effects. A final recommendation in the work above is simply to conduct more demand-treatment experiments. We recommend including SWB outcomes in such future studies.

Of course, if SWB survey respondents do not know that they are participating in an experiment, there is little scope for demand effects. When they know they are participants but do not know what other treatments (or controls) exist, there is some scope for demand effects, but it is still reduced relative to experiments with fully transparent experimental design.³¹ In contrast, transparent design may help reduce treatment-dependent *context* effects, to the extent that it provides all participants with a similar context; while withholding information regarding other treatments may mean that participants in different treatments are subject to very different contexts. Thus, as in other cases, trying to minimize both demand and context effects may involve experimental-design tradeoffs.

We close this section with our repeated recommendation from earlier in this chapter, i.e., to routinely collect, in relevant experiments, a variety of SWB measures. In the context of potential demand effects, this catchall recommendation has two additional potential advantages. First, following Ifcher and Zarghamee (2011), we recommend adding also SWB aspects that are not expected *ex ante* to be affected differently across treatments. Referred

³¹de Quidt et al. (2019) make a related point, focusing on the distinction between within- and between-subject designs:

With respect to experimenter demand, the advantage of a between-subject design is that each participant is only exposed to one treatment environment. . . . Without direct knowledge of the treatment variable, between-subject designs make it difficult for participants to guess what, if any, parts of the experimental environment are being varied between treatments. For example, is it the structure or scale of incentives? The framing? The number of other game participants? And in what “direction” does the variation occur? Are they in the high-stakes or low-stakes treatment?

We note that while in within-subject designs, the manipulation is indeed typically transparent, in between-subjects design it is not always opaque. Falk and Graeber (2020), for example, use a between-subjects design with transparent assignment (using a lottery). Moreover, this very transparency is a fundamental design element, as each of the two €100-received-versus-€350-donated conditions includes, as a crucial part of the treatment, *not* begin assigned into the opposite condition. We also note that for SWB in large field interventions, obfuscated follow-ups may be more relevant than an opaque between-subjects design. If the same research team that offered participants large money transfers or housing vouchers is later on asking them SWB questions, they may correctly infer that the team hopes for improved SWB. Control participants, in contrast, may not infer that improved SWB in hoped for. Importantly, both groups may do so even without said “direct knowledge of the treatment variable.”

to above as placebo measures, they could indeed also serve as demand-effect falsification tests.

Second, following the demand-treatments literature, we also recommend going in the opposite direction, i.e., adding SWB questions that are expected *ex ante*, by design, to be more subject to demand effects. We are not aware of research attempting to identify such “anti-placebos” so, at present, we do not know how to fully implement this recommendation. We do know however that demand effects are expected to affect more than one outcome in the same direction, and that collecting multiple outcomes will allow researchers to assess this prediction. As discussed above, Haushofer and Shapiro (2016) use such reasoning to argue against a demand-effect explanation. While their argument is conveyed as *ex post*, future researchers could purposefully include a variety of SWB measures also for that purpose.

While we do not have direct evidence on demand effects on SWB, their assessment, or their mitigation, collecting a variety of different SWB measures may help make progress on these questions too.

6 Conclusion

The goal of this chapter is to encourage researchers to carefully and thoughtfully collect SWB data in their experiments. We did not discuss how to analyze these data—a topic covered by a huge literature on ordered survey responses (including recent work on scale use). Neither did we warn about an inherent tradeoff: Collecting many measures increases the opportunity for both true and false discoveries—a topic covered by a huge literature on multiple hypothesis testing. We do remind readers that the interpretation of almost any analysis of SWB data—including, e.g., the most standard comparison of mean SWB responses across experimental conditions—relies on a list of necessary assumptions. As recently discussed in a review by Benjamin et al. (2024), some of these are strong assumptions; some are untestable; some have not been widely tested; and some have been documented not to hold for widely used SWB measures, in important settings, and in common applications. We thus also call for careful *interpretation* of SWB results.

That recent review closes with a table (Table 2, p. 385) that organizes the authors’ advice by two dimensions of cost. We close this chapter with Table 1, our own version of an advice table. Like that table, on the vertical dimension we consider costs in survey design, implementation, effort, and duration. Here, we ask: Does our advice involve incurring additional such costs? On the horizontal dimension we indicate whether we think that reliable advice is already available or requires further research and development. The two tables are complementary: There is little overlap in the content of advice across them, as

the review by Benjamin et al. (2024) and this chapter cover different topics. In particular, that review has a lot to say about interpreting SWB vis-à-vis preferences—a topic that this chapter does not touch, but that economists would be remiss to ignore. More generally, the tables’ similar structure is meant to encourage economists using SWB data to consult both and, we hope, to add their own advice to the two tables over time.³²

Both our parsing of advice into concrete pieces and our assignment of each piece into a specific quadrant in Table 1 could be questioned; they are meant to be merely illustrative and, we hope, useful. The “Readily apply” column includes both low-cost actions such as being clear about what SWB data are hoped to capture and how terms are used (top-left quadrant) and more costly actions involving collecting more data and employing more expensive study designs (bottom-left). Depending on their resources, researchers can already implement these recommendations.

On the other hand, as the “Research & develop” column makes clear, at present we do not have ready-made advice on some potentially important concerns. We provide what we believe, based on existing research discussed in this chapter, are promising directions for further investigation. Rather than discouraging researchers from collecting SWB data, we hope that this column would have the opposite effect of encouraging researchers to collect *additional* SWB data that would also help make progress on these issues. Indeed, besides hoping to help researchers include SWB outcomes in their own present studies, this chapter’s data-collection focus is meant to serve an additional purpose: If successful, a joint, decentralized data-collection effort would also mean that over time, SWB data get explored, accumulated, and, hopefully, better understood.

References

- Adler, M. D. (2019). *Measuring Social Welfare: An Introduction*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/oso/9780190643027.001.0001>
- Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The Welfare Effects of Social Media. *American Economic Review*, 110(3), 629–676. <https://doi.org/10.1257/aer.20190658>

³²Beyond several important warnings regarding the interpretation of SWB vis-à-vis preferences, advice in Table 2 in Benjamin et al. (2024) regarding topics not covered in this chapter also includes: using SWB questions with (quasi-)continuous response scales; considering the addition of calibration questions for assessing, and correcting for, differences in individual response-scale use; using introspective questions after SWB questions to assess what respondents interpreted the SWB question to be asking about; and eliciting tradeoff-based weights for aggregating different SWB domains into a general well-being index.

Table 1: Practical Advice for SWB Data Collection, by Cost and Applicability

		Readily apply or further research & develop?	
		Readily apply	Research & develop
Additional survey design, implementation, effort, or duration costs?	No	Ask yourself (and state) what you are trying to measure (2.2) Clarify the SWB terms used (3.2) Beware of comparisons across demographic groups (3.1) Beware of “pricing” regressions (3.2) Randomize question order (3.1, 4.3) Use buffer/transition questions (4.1)	Use specific language about time horizon & social circles (2.2, 3.2, 4.3) Beware of context and demand effects (3, 4, 5)
	Yes	Collect multi-item measures (2.1, 3) Measure multiple times (2.1, 3) <u>Lab:</u> Compare SWB vs. choice (3.1) <u>Field:</u> Collect a variety of affective, evaluative, and eudaimonic measures, over different time horizons, at different frequencies (3.2) Apply best-practice advice for demand-effect mitigation (5.3)	Include “placebo SWB measures” (3) Include demand-effects “anti-placebos” (5.3) Communicate question intention (4.2, 4.3) Minimize demand & context effects through study design, design transparency, & data collection (3, 4.3, 5.3)

Notes: In parentheses, sections where advice is discussed and/or demonstrated.

- Allcott, H., Gentzkow, M., & Song, L. (2022). Digital Addiction. *American Economic Review*, 112(7), 2424–2463. <https://doi.org/10.1257/aer.20210867>
- Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484), 1481–1495. <https://doi.org/10.1198/016214508000000841>
- Benjamin, D. J., Cooper, K., Heffetz, O., & Kimball, M. (2024). From Happiness Data to Economic Conclusions. *Annual Review of Economics*, 16, 359–391. <https://doi.org/https://doi.org/10.1146/annurev-economics-081623-021136>
- Benjamin, D. J., Cooper, K., Heffetz, O., Kimball, M. S., & Zhou, J. (2023). *Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being* (Working Paper No. 31728). National Bureau of Economic Research. <https://doi.org/10.3386/w31728>
- Benjamin, D. J., Debnam Guzman, J., Fleurbaey, M., Heffetz, O., & Kimball, M. S. (2023). What do Happiness Data Mean? Theory and Survey Evidence. *Journal of the European Economic Association*, 21(6), 2377–2412. <https://doi.org/10.1093/jeea/jvad026>
- Benjamin, D. J., Heffetz, O., Kimball, M. S., & Rees-Jones, A. (2012). What Do You Think Would Make You Happier? What Do You Think You Would Choose? *American Economic Review*, 102(5), 2083–2110. <https://doi.org/10.1257/aer.102.5.2083>
- Benjamin, D. J., Heffetz, O., Kimball, M. S., & Rees-Jones, A. (2014). Can Marginal Rates of Substitution Be Inferred from Happiness Data? Evidence from Residency Choices. *American Economic Review*, 104(11), 3498–3528. <https://doi.org/10.1257/aer.104.11.3498>
- Benjamin, D. J., Heffetz, O., Kimball, M. S., & Szembrot, N. (2014). Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. *American Economic Review*, 104(9), 2698–2735. <https://doi.org/10.1257/aer.104.9.2698>
- Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. London: T. Payne; Son.
- Bernheim, B. D., Kim, K., & Taubinsky, D. (2024). *Welfare and the Act of Choosing* (Working Paper No. 32200). National Bureau of Economic Research. <https://doi.org/10.3386/w32200>
- Bischoff, I., & Frank, B. (2011). Good news for experimenters: Subjects are hard to influence by instructors' cues. *Economics Bulletin*, 31(4), 3221–3225. <https://ideas.repec.org/a/ebl/ecbull/eb-11-00779.html>
- Bishop, G. F. (1987). Context Effects on Self-Perceptions of Interest in Government and Public Affairs. In H. J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information*

- Processing and Survey Methodology* (pp. 179–199). New York: Springer. https://doi.org/10.1007/978-1-4612-4798-2_10
- Brañas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology*, 28(4), 477–486. <https://ideas.repec.org/a/eee/joepsy/v28y2007i4p477-486.html>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436. <https://doi.org/10.1126/science.aaf0918>
- Cantril, H. (1965). *The pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.
- Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review*, 102(6), 2981–3003. <https://doi.org/10.1257/aer.102.6.2981>
- Cilliers, J., Dube, O., & Siddiqi, B. (2015). The white-man effect: How foreigner presence affects behavior in experiments. *Journal of Economic Behavior & Organization*, 118, 397–414. <https://doi.org/https://doi.org/10.1016/j.jebo.2015.03.015>
- Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, 24, 385–396. <http://www.jstor.org/stable/2136404>
- Deaton, A. (2012). The financial crisis and the well-being of Americans: 2011 OEP Hicks Lecture. *Oxford Economic Papers*, 64(1), 1–26. <https://doi.org/10.1093/oep/gpr051>
- Deaton, A., & Stone, A. A. (2016). Understanding context effects for a measure of life evaluation: how responses matter. *Oxford Economic Papers*, 68(4), 861–870. <https://doi.org/10.1093/oep/gpw022>
- de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and Bounding Experimenter Demand. *American Economic Review*, 108(11), 3266–3302. <https://doi.org/10.1257/aer.20171330>
- de Quidt, J., Vesterlund, L., & Wilson, A. J. (2019). Experimenter Demand Effects. In A. Schram & A. Ule (Eds.), *Handbook of Research Methods and Applications in Experimental Economics* (pp. 384–400). Northampton, MA: Edward Elgar Publishing. <https://doi.org/https://doi.org/10.4337/9781788110563.00031>
- de Quidt, J., Vesterlund, L., & Wilson, A. J. (2024). Experimenter Demand Effects. In A. Rees-Jones (Ed.), *Handbook of Experimental Methods in the Social Sciences*. Edward Elgar, forthcoming.

- Devoto, F., Duflo, E., Dupas, P., Parienté, W., & Pons, V. (2012). Happiness on Tap: Piped Water Adoption in Urban Morocco. *American Economic Journal: Economic Policy*, 4(4), 68–99. <https://doi.org/10.1257/pol.4.4.68>
- Diaz, L., Houser, D., Ifcher, J., & Zarghamee, H. (2023). Estimating social preferences using stated satisfaction: Novel support for inequity aversion. *European Economic Review*, 155, 104436. <https://doi.org/https://doi.org/10.1016/j.eurocorev.2023.104436>
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, 49(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13
- Diener, E., Oishi, S., & Tay, L. (Eds.). (2018). *Handbook of Well-Being*. Salt Lake City, UT: DEF Publishers.
- Dolan, P., & Metcalfe, R. (2012). Measuring Subjective Wellbeing: Recommendations on Measures for use by National Governments. *Journal of Social Policy*, 41(2), 409–427. <https://doi.org/10.1017/S0047279411000833>
- Easterlin, R. A. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In P. A. David & M. W. Reder (Eds.), *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz* (pp. 89–125). Cambridge, MA: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-205050-3.50008-7>
- Falk, A., & Graeber, T. (2020). Delayed negative effects of prosocial spending on happiness. *Proceedings of the National Academy of Sciences*, 117(12), 6463–6468. <https://doi.org/10.1073/pnas.1914324117>
- Fumagalli, E., & Fumagalli, L. (2022). Subjective well-being and the gender composition of the reference group: Evidence from a survey experiment. *Journal of Economic Behavior & Organization*, 194, 196–219. <https://doi.org/https://doi.org/10.1016/j.jebo.2021.12.016>
- Haaland, I., & Roth, C. (2020). Labor market concerns and support for immigration. *Journal of Public Economics*, 191, 104256. <https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104256>
- Haaland, I., & Roth, C. (2023). Beliefs about Racial Discrimination and Support for Pro-Black Policies. *The Review of Economics and Statistics*, 105(1), 40–53. https://doi.org/10.1162/rest_a_01036
- Haaland, I., Roth, C., & Wohlfart, J. (2023). Designing Information Provision Experiments. *Journal of Economic Literature*, 61(1), 3–40. <https://doi.org/10.1257/jel.20211658>
- Haushofer, J., & Shapiro, J. (2016). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *The Quarterly Journal of Economics*, 131(4), 1973–2042. <https://doi.org/10.1093/qje/qjw025>

- Heffetz, O., & Rabin, M. (2013). Conclusions Regarding Cross-Group Differences in Happiness Depend on Difficulty of Reaching Respondents. *American Economic Review*, 103(7), 3001–3021. <https://doi.org/10.1257/aer.103.7.3001>
- Hughes, M. E., Waite, L. J., Hawkey, L. C., & Cacioppo, J. T. (2004). A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. *Research on Aging*, 26(6), 655–672. <https://doi.org/10.1177/0164027504268574>
- Huppert, F. A., Marks, N., Clark, A., et al. (2009). Measuring Well-being Across Europe: Description of the ESS Well-being Module and Preliminary Findings. *Social Indicators Research*, 91, 301–315. <https://doi.org/10.1007/s11205-008-9346-0>
- Ifcher, J., & Zarghamee, H. (2011). Happiness and Time Preference: The Effect of Positive Affect in a Random-Assignment Experiment. *American Economic Review*, 101(7), 3109–3129. <https://doi.org/10.1257/aer.101.7.3109>
- Ifcher, J., Zarghamee, H., & Goff, S. H. (2022). Happiness in the Lab: What Can Be Learned About Subjective Well-Being from Experiments? In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics* (pp. 1–27). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-57365-6_187-1
- Ifcher, J., Zarghamee, H., Houser, D., & Diaz, L. (2020). The relative income effect: An experiment. *Experimental Economics*, 23, 1205–1234. <https://doi.org/10.1007/s10683-020-09648-w>
- Kahneman, D., Diener, E., & Schwarz, N. (Eds.). (1999). *Well-being: The foundations of hedonic psychology*. New York: Russell Sage Foundation. <http://www.jstor.org/stable/10.7758/9781610443258>
- Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of Experienced Utility. *The Quarterly Journal of Economics*, 112(2), 375–406. <https://doi.org/10.1162/003355397555235>
- Kőszegi, B., & Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92(8–9), 1821–1832. <https://doi.org/10.1016/j.jpubeco.2008.03.010>
- Layard, R. (2005). *Happiness: Lessons from a New Science*. London: Penguin.
- Lucas, R. E. (2018). Reevaluating the strengths and weaknesses of self-report measures of subjective well-being. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being*. Salt Lake City, UT: DEF Publishers.
- Lucas, R. E., Oishi, S., & Diener, E. (2016). What we know about context effects in self-report surveys of well-being: comment on Deaton and Stone. *Oxford Economic Papers*, 68(4), 871–876. <https://doi.org/10.1093/oeq/gpw023>

- Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., et al. (2012). Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science*, 337, 1505–1510. <https://doi.org/10.1126/science.1224648>
- Lyubomirsky, S., & Lepper, H. S. (1999). A Measure of Subjective Happiness: Preliminary Reliability and Construct Validation. *Social Indicators Research*, 46(2), 137–155. <https://doi.org/10.1023/A:1006824100041>
- McBride, M. (2010). Money, happiness, and aspirations: An experimental study. *Journal of Economic Behavior & Organization*, 74(3), 262–276. <https://doi.org/https://doi.org/10.1016/j.jebo.2010.03.002>
- Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4), 371–378. <https://doi.org/10.1037/h0040525>
- Mummolo, J., & Peterson, E. (2019). Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review*, 113(2), 517–529. <https://doi.org/10.1017/S0003055418000837>
- Nussbaum, M. C. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge, UK: Cambridge University Press.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Pavot, W. (2018). The Cornerstone of Research on Subjective Well-Being: Valid Assessment Methodology. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of Well-Being* (pp. 83–93). Salt Lake City, UT: DEF Publishers. <https://doi.org/nobascholar.com>
- Peterson, R. A., & Sauber, M. (1983). A mood scale for survey research. In P. E. Murphy et al. (Eds.), *American marketing association educators' proceedings* (pp. 409–414). Chicago: American Marketing Associations.
- Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, 1(3), 385–401. <https://doi.org/10.1177/014662167700100306>
- Robinson, M. D., & Klein, R. J. (2018). What do subjective well-being judgments mean? Sources and distinctions, processes and mechanisms. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being*. Salt Lake City, UT: DEF Publishers.
- Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, 89(3), 395–406. <https://doi.org/10.1037/0022-3514.89.3.395>
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93–105. <https://doi.org/10.1037/0003-066X.54.2.93>

- Schwarz, N., & Strack, F. (1999). Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology* (pp. 61–84). New York: Russell Sage Foundation.
- Sen, A. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.
- Stantcheva, S. (2023). How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. *Annual Review of Economics*, 15, 205–234. <https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157>
- Stone, A. A., Walentynowicz, M., Schneider, S., Junghaenel, D. U., Broderick, J. E., & Deaton, A. (2022). Item Context Effects Are Relevant for Monitoring Evaluative Well-being: Replication of Previous Work and Mitigation. *Field Methods*, 34(1), 36–51. <https://doi.org/10.1177/1525822X221077393>
- Tourangeau, R. (1992). Context Effects on Responses to Attitude Questions: Attitudes as Memory Structures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 35–47). New York, NY: Springer. https://doi.org/10.1007/978-1-4612-2848-6_4
- Tsutsui, K., & Zizzo, D. J. (2014). Group status, minorities and trust. *Experimental Economics*, 17, 215–244. <https://doi.org/10.1007/s10683-013-9364-x>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., et al. (2017). The Effect of Mood on Judgments of Subjective Well-Being: Nine Tests of the Judgment Model. *Journal of Personality and Social Psychology*, 113(6), 939–961. <https://doi.org/10.1037/pspp0000115>
- Zizzo, D. J. (2010). Experimenter Demand Effects in Economic Experiments. *Experimental Economics*, 13, 75–98. <https://doi.org/10.1007/s10683-009-9230-z>