Vienna Yearbook of Population Research 2021 (Vol. 19), pp. 67–89

Comparing global reports of subjective well-being to experiential measures

Richard E. Lucas[1,*]

Abstract

Subjective well-being (SWB) is an overall evaluation of the quality of a person's life from his or her own perspective. One common method of assessing this construct requires respondents to think about their life as a whole and to provide a "global" evaluation that summarizes across life domains or affective experiences over extended periods of time. The validity of these global measures has been challenged, however; and experiential measures, which ask respondents to report on their momentary evaluative experiences many times over a constrained time period, have been suggested as a more valid alternative. This paper addresses the empirical evidence for one important challenge to global measures: the possibility that temporarily salient information overwhelmingly influences global judgments, reducing their reliability and validity. This paper critiques prior evidence for this challenge and presents new concerns about the assumed validity of the proposed alternative: experiential measures.

Keywords: subjective well-being; life satisfaction; measurement; experience sampling method; day reconstruction method

1 Introduction

Subjective well-being (SWB) is an overall evaluation of the quality of a person's life from his or her own perspective. According to Diener et al. (1999), the construct and its measures are characterized by three features. First, SWB, as its name suggests, is *subjective*; it captures the respondent's own perspective about how life is going rather than an objective evaluation or an outside observer's perspective. Second, it is *global* in scope. In other words, measures of SWB are intended to capture overall evaluations that incorporate multiple domains of a person's life. Finally, SWB is

[1]Department of Psychology, Michigan State University, East Lansing, Michigan, USA

[*]Correspondence to: Richard E. Lucas, lucasri@msu.edu

DOI: 10.1553/populationyearbook2021.rev01

balanced in its focus, capturing both positive and negative aspects of life. These distinguishing features of SWB make it especially well-suited for use in a variety of applied and theoretical investigations. For instance, by focusing on subjective evaluations of life as a whole, SWB researchers allow respondents to consider, evaluate and weight the importance of a broad range of factors that may contribute to their perceived quality of life, rather than having those factors selected for them. The subjective nature of the construct contrasts with alternative approaches that rely on expert judgments or theoretical arguments to determine which features of the good life are most important (e.g., Hurka 2014; Ryan and Deci 2000; Ryff 1989).

The subjectivity of SWB and its measures are especially important when investigating decisions, events or interventions that may affect some life domains in a positive manner, while simultaneously affecting other life domains negatively. For instance, specific medical treatments might successfully reduce pain or other symptoms, while also causing negative side effects for cognitive or social functioning. Thus, SWB may be useful if the aim is to find out whether, on balance, the treatment has improved a respondent's life. The global, subjective nature of SWB judgments allows respondents to evaluate and weight the impact of the treatment on these and other life domains. By assessing SWB, researchers can more fully evaluate the total effect of treatments or other interventions. These features have allowed SWB measures to be incorporated into a wide range of applied and theoretical research programs, and they have led to an increased focus on SWB judgments in policy decisions (Diener et al. 2009; Dolan and White 2007; Frijters and Krekel 2021; Kahneman et al. 2004b; Krueger et al. 2009).

Of course, the usefulness of SWB as a construct depends critically on the quality of the measures that are available to assess it, and debates about measurement issues have existed for as long as SWB has been studied. The greatest strength of SWB—its subjective nature—is also responsible for its greatest measurement challenge: SWB cannot be directly observed, and no "gold standard" measure of the construct exists.[1] Thus, substantive investigations of SWB must simultaneously grapple with standard issues of internal and external validity, and with ever-present concerns about the quality of the measures used to assess the central construct.

Throughout much of the early history of research on SWB, social scientists relied on *global evaluative* measures. This class of measures requires respondents to reflect on their lives and to provide an overall evaluation. Examples include the Satisfaction With Life Scale (Diener et al. 1985), along with a range of single-item life satisfaction scales used in many population-based surveys. Although global evaluative measures have a number of advantages, including their simplicity and high face validity, they are, of course, not perfect. One important challenge to these

[1] One might argue that subjective evaluations should correspond to internal, physiological states, and thus, that physiological measures could ultimately serve as a gold standard. Although this is a possible outcome as psychologists' understanding of physiological indicators develops, there are currently no candidate measures that have been found to have substantial levels of validity as measures of the underlying construct (Ito and Cacioppo 1999).

measures—a challenge that is one primary focus of this paper—has been proposed by researchers from the judgment model perspective (Schwarz and Strack 1999). Proponents of the judgment model suggest that constructing a global evaluation should be time-consuming and difficult; i.e., it should require respondents to remember, evaluate and aggregate evaluations of a potentially infinite set of life domains and events. As a result, respondents are thought to rely on a variety of heuristics and biases that simplify the evaluation process, but also affect the validity of the measures that are found.

If these judgment processes affect the reliability and the validity of global measures, then it would be fruitful to investigate alternative measurement strategies that address these specific concerns. Indeed, global measures can be contrasted with *experiential* measures, which focus on narrower evaluations of specific moments in time (Kahneman and Riis 2005). For instance, with experience sampling methods, respondents are asked repeatedly over an extended period to provide ratings of how they feel at each moment. These momentary ratings can capture emotional reactions to specific events, or they can capture longer-lasting moods that may be influenced by an individual's temperament or lingering effects of events or circumstances that are not currently the focus of attention. Presumably, people whose lives are going well will evaluate more of the moments of their lives positively (Kahneman 1999). Importantly, from a measurement perspective, experiential measures should be easier for respondents to complete because respondents are not required to remember and aggregate across multiple moments or life domains. In other words, experiential measures might solve the problems of global reports because they simplify the judgment that is required: i.e., respondents only need to consider the moment that they are in.

The goals of this paper are twofold. First, I critically review the evidence that supports the judgment model critique of global evaluative measures. Although a number of highly cited studies purport to provide evidence for the idea that temporarily salient information can dramatically impact global judgments of SWB, these studies have many characteristics that limit the strength of the evidence they provide. The second goal of the paper is to evaluate the psychometric properties of experiential measures, which have been proposed as a useful alternative to global measures. I will argue that although these measures directly address the concerns about global measures, they may pose their own threats to validity that have previously not been considered.

1.1 Measurement challenges: The judgment model of subjective well-being

One of the clearest challenges to the validity of self-reports of SWB comes from Schwarz and Strack's *judgment model* of SWB (for a review, see Schwarz and Strack 1999). The judgment model focuses on global evaluative measures of subjective well-being: i.e., measures like life satisfaction judgments, which ask respondents

to consider their life as a whole and to provide a summary evaluation. According to the judgment model, responses to these global measures result from a judgment process that is susceptible to systematic sources of bias and invalidity. Multiple studies from the judgment model perspective have investigated specific processes that are purported to occur when these judgments are made.

For instance, one concern is that SWB judgments are made too quickly to reflect an overall evaluation of all relevant information about a person's life. Thus, people may quickly scan their memories (or even their current thoughts and feelings) for relevant information, relying too heavily on the information—including potentially unimportant information—that happens to be accessible. Indeed, Schwarz and Strack (1999) suggested that "information that has just been used—for example, to answer a preceding question in a questionnaire—is particularly likely to come to mind later on" (p. 63), and that this temporarily accessible information biases the resulting responses.

In a famous demonstration of this effect, Strack et al. (1988) randomly assigned respondents to answer two questions, one about their global life satisfaction and one about their satisfaction with dating, in one of two orders: dating satisfaction first or global satisfaction first. According to the judgment model, when dating satisfaction is presented first, this life domain should be made especially salient, which should result in strong correlations between the two questions. In contrast, when the global evaluation is presented first, the respondent's romantic life may not be on his or her mind, and may thus have less of an impact. Results from two separate studies supported this prediction: Correlations between the two questions were only .16 and −.12 when the global satisfaction question was asked first, but they were .55 and .66 when the satisfaction with dating question was asked first. Thus, these studies suggest that the information that happens to be accessible at the time of a global judgment affects the judgment itself.[2]

Importantly, it is not just information about a person's life that can affect these judgments; according to judgment model researchers, even feelings that people are experiencing at the time of judgment can affect their judgments in undesirable ways. In one of the most cited examples of research from this tradition, Schwarz and Clore (1983) conducted two studies to assess whether respondents tended to rely on potentially irrelevant current moods when making judgments about their life as a whole. In each of these studies, the authors manipulated the respondents' current mood (using recall of positive and negative life events in one study and changes in daily weather in the second) to assess whether these manipulations also affected their life satisfaction judgments. In both studies, the respondents in the positive mood conditions reported substantially higher life satisfaction than those

[2] Deaton and Stone (2016) describe an unexpected item-order effect in a large sample of Americans that could be interpreted as further support for this type of context effect. However, because the study was not developed specifically to elicit this effect, the interpretation of this effect is somewhat ambiguous, and more research into the processes that underlie it is needed (Lucas et al. 2016).

in the negative mood conditions (with effect sizes close to a full standard deviation difference). These compelling results were also replicated in other studies that used different mood induction procedures, including placing respondents in a pleasant or an unpleasant environment (Schwarz et al. 1987), surveying people before and after important soccer games with varying outcomes (Schwarz et al. 1987), and even simply arranging for respondents to find a dime before completing the relevant survey (Schwarz 1987).

The findings of other studies have even suggested that feelings that arise simply as a result of the process of searching one's memory for relevant information can affect judgments. For instance, although this study was not conducted with well-being judgments as an outcome, Schwarz et al. (1991) found that the ease with which respondents could recall examples relevant to a characteristic could affect their ratings of that characteristic. Specifically, participants who were asked to list six examples of times when they acted assertively rated themselves as more assertive than those who were asked to list 12 examples of assertive behaviors. Presumably, the feelings respondents had about the difficulty associated with producing 12 examples provided information that was used in the judgment itself. To my knowledge, studies of this accessibility effect have not been conducted in the well-being domain. However, evidence of this general phenomenon adds support to the idea that a variety of feelings that people experience (including those that are arguably irrelevant for the judgment at hand) can influence their broad judgments.

Taken together, these studies suggest that irrelevant information that happens to be on a person's mind at the time of judgment can have strong effects on the resulting judgment. These findings represent an important challenge to the validity of global reports of SWB because they suggest that people do not always make SWB judgments in a manner that is consistent with theory or intuition, and that irrelevant information may play a large role in their responses. Indeed, after reviewing research from the judgment model tradition, Schwarz and Strack suggested that one plausible conclusion could be that "There is little to be learned from global self-reports of well-being ... [What] is being assessed, and how, seems too context dependent to provide reliable information about a population's well-being, let alone information that can guide public policy" (1999, p. 80). These authors argued that the instability inherent in the judgment process limited the extent to which SWB judgments reflected the actual quality of the respondents' lives. Research from the judgment model tradition is often cited in critiques of well-being research, and the presumed problems that these judgment processes create for the validity of these measures have led some scholars to suggest that alternatives to global measures of well-being are needed (e.g., Kahneman 1999; National Academy of Sciences 2013; OECD 2013; Testoni et al. 2018).

2 The promise of experiential measures

The concerns that have been raised about global evaluative measures have led some researchers to propose and investigate alternative approaches for assessing SWB. Many of these approaches have focused on addressing the possibility that respondents may have difficulties answering questions that call for global judgments because answering such questions requires respondents to call to mind information from a broad range of life domains, and to aggregate that information into a single judgment. Both tasks—remembering and aggregating—may be challenging and subject to biases.

Robinson and Clore (2002) proposed an accessibility model focused on self-reports of emotion that is relevant (and that has been applied) to judgments of SWB. In their model, they proposed that different time frames for emotion reports result in the use of different processes. For instance, they argued, when evaluating current emotional experiences, people may rely on experiential knowledge, which is likely to capture actual variance in emotional experiences. At the very least, they pointed out, reports of current emotions cannot be affected by flawed memory, which suggests that these reports should be relatively accurate.

As the time frame of the emotional report increases in duration (the amount of time it covers) or temporal distance (the amount of time that has elapsed since the emotion occurred), respondents must rely on episodic memory—i.e., memory of a specific episode and the details of the experience—to provide a report. This process requires the respondent to reconstruct the episode, and memory limitations may prevent respondents from accurately recalling how they felt in the moment. Furthermore, biases related to salient features of specific experiences may systematically distort their recall of emotional experiences. Thus, it might be expected that reports of recent emotional experiences are somewhat accurate, but less so than emotions reported in the moment.

As the length of time between the experience and the report grows, the details of the experience become more difficult to remember, and, according to Robinson and Clore (2002), individuals may rely even more heavily on situation-specific beliefs to reconstruct their experienced emotions. Indeed, respondents may rely more on semantic knowledge about the nature of events than on episodic memory for a specific event when reconstructing the emotional episode. If general expectations about situations contradict the unique features of a specific experience, the reports of that experience may be inaccurate. For instance, if a respondent believes that she typically experiences joy at parties and forgets that during a particular party one month previously she was concerned about a work-related conflict, she may overestimate the level of joy she experienced. Finally, Robinson and Clore (2002) suggested, if respondents forget what they were doing during the time period in question, then they may rely on general beliefs about how they typically feel, even if their experience had actually deviated from these typical levels.

Robinson and Clore's (2002) model raises concerns about people's ability to remember relevant information when providing responses to global well-being

questions. However, there are also concerns about people's ability to aggregate across experiences even if their memory was flawless. For instance, in a series of studies, Kahneman and colleagues provided evidence that when asked to evaluate an extended experience that varied in hedonic valence, respondents tended to place less emphasis on the duration of the experience, and to weigh the peak experience and the end experience heavily when providing an overall evaluation (Fredrickson and Kahneman 1993; Kahneman et al. 1993; Redelmeier and Kahneman 1996). In a particularly striking demonstration of this effect, Redelmeier et al. (2003) randomly assigned patients undergoing colonoscopies to either have the procedure end normally after an especially painful part of the procedure, or to have the procedure extended with a short interval of much milder pain. The authors predicted that even though the patients in the extended condition experienced more pain overall, the tendency to neglect the duration of an experience, combined with the greater salience of the much milder pain at the end of an episode, would lead to more positive evaluations of the extended procedure than of the normal procedure. In support of this prediction, the participants in the extended procedure condition evaluated it more positively, and were even more likely to return for a follow-up than those in the normal procedure condition (although the latter effect was small and just barely significant). Because the two procedures were, on average, identical other than the added interval of mild pain, the tendency to evaluate the longer experience as more desirable than the shorter one violated reasonable expectations about which experience was objectively better. Although debates continue about the processes that underlie effects like these (e.g., Tully and Meyvis 2016), there are plausible concerns about the extent to which people can accurately aggregate across experiences.

Due in part to these concerns, researchers turned to procedures for assessing well-being that reduce reliance on memory and the need for aggregation by respondents. Specifically, *experiential* measures of SWB shift the emphasis from people's evaluations of their life as a whole to their evaluations of the individual experiences that they have (Kahneman and Riis 2005). The potential value of these approaches is highlighted most clearly by Kahneman (1999) in a chapter titled "Objective Happiness." In this chapter, Kahneman argued that something close to an objective measure of SWB can be obtained by repeatedly asking respondents a very simple question that should be unaffected by the concerns raised above: namely, whether, at the current moment, their experience is positive or negative. These momentary evaluations can then be aggregated to get an assessment of the respondent's overall experience.

Although Kahneman (1999) focused on a very specific type of question for his proposed measure of objective happiness (the simple dichotomous evaluation of whether the person was feeling positive or negative), a broad range of experiential measures have been examined as alternatives to global evaluative measures. There is a long history of research that samples people's experiences repeatedly over time, while focusing on the emotions and feelings that people report in those moments (for an overview, see Mehl and Conner 2013). These momentary reports can be

aggregated to get an overall sense of the positivity of people's experiences, which could be considered to be a reasonable proxy for an evaluation of their life as a whole (for a discussion of whether such measures should be seen as equivalent, see Diener et al. 2018).

Of course, methods that repeatedly sample experiences over time are quite burdensome both for participants and researchers. To tackle this problem, researchers have developed alternative procedures that streamline assessments while attempting to generate a rich set of data over time. Specifically, Kahneman et al. (2004a) proposed the *day reconstruction method*, in which respondents reconstruct an entire day at a time, listing distinct episodes that they experienced during the day, along with the moods and feelings that they experienced during those episodes. The goal of the day reconstruction method is to obtain rich data about momentary experiences using techniques that are less burdensome than experience sampling, but that do not have the same problems as global measures of SWB. Because the day reconstruction method can be administered in a single assessment, it has been incorporated into several large-scale studies, including the German Socio-Economic Panel Study (Goebel et al. 2019), the Panel Study of Income Dynamics,[3] and the American Time Use Survey.[4] This increasingly widespread adoption of experiential measures shows that they are a promising complement or alternative to traditional global measures.

3 Comparing global reports to experiential measures

If the validity of global reports of SWB is severely affected by the processes described by judgment model researchers, then experiential measures provide a sensible alternative that solves the worst of these problems. Specifically, if the problem with global reports is that people are unable or unwilling to consider all relevant information when making global judgments, and instead rely on whatever information happens to be accessible, experiential measures solve the problem by only asking respondents to consider a narrow moment of time, as that moment occurs. Alternatively, if the primary problem is that people focus on especially salient moments or features when aggregating across a range of experiences, experiential measures can solve this problem by repeatedly assessing moments, while taking the task of aggregation out of the hands of the respondents.

In short, researchers have argued for the benefits (if not the superiority; cf. Kahneman 1999) of using experiential measures instead of global measures precisely because the former directly address concerns that have been raised about the latter. However, it is possible to question whether the evidence that the judgment model researchers have presented really does robustly challenge the validity of global report measures. In the next section, I argue that the evidence for the existence of

[3] https://psidonline.isr.umich.edu/data/Documentation/UserGuide2017.pdf

[4] https://www.bls.gov/tus/atususersguide.pdf

such problems has been overstated. In addition, because experiential measures so clearly solve two problems associated with global measures (problems of memory and aggregation), there has been a sense of complacency about the validity of these experiential measures. Thus, few explicit tests of their validity have been performed. Indeed, I argue that this complacency has prevented researchers from considering whether experiential measures may pose their own unique challenges to validity that do not affect global measures. If this is the case, then the important questions about the relative strengths of the two types of measures of SWB would remain.

3.1 Concerns about the judgment model

Over the past decade, scientists have been grappling with concerns about the replicability of previously published work (e.g., Open Science Collaboration 2015). Although concerns about replicability have touched many areas of science, the field of social psychology—the field that encompasses judgment model research—has been at the heart of this debate. Social psychological studies fared poorly in the first major attempt to replicate large sets of high-profile studies (Open Science Collaboration 2015), and more targeted investigations of specific high-profile findings have also repeatedly failed to replicate initial claims (e.g., O'Donnell et al. 2018; Cheung et al. 2016; Eerland et al. 2016; McCarthy et al. 2018). To be sure, researchers do not know the true rate of replicability of social psychological studies, but the substantial number of failed replications that has emerged has led to a renewed skepticism about even the most foundational studies in the field.

Specific concerns about problematic research practices have strengthened this skepticism. For instance, researchers (not just from social psychology) have admitted to a variety of such practices, including optional stopping (i.e., checking the results before the data collection is complete and stopping once significance is obtained); the strategic omission of outliers, measures, or even experimental conditions that do not support hypotheses; and post hoc theorizing that makes it appear as though unpredicted effects were hypothesized (John et al. 2012; Kerr 1998). When combined with a tendency to use underpowered designs (Button et al. 2013) and a broad bias against publishing null results, these practices can lead to a high rate of false positives in the literature (Ioannidis 2005). Indeed, in a famous investigation of the consequences of these practices, Simmons et al. (2011) showed that support for clearly wrong (and nonsensical) hypotheses can easily be found when a combination of these practices is used.

These issues are important for discussions about the measurement of SWB. When highly cited studies from the judgment model are evaluated from this modern, skeptical perspective, there are reasons for concern. Specifically, almost all original effects in support of the judgment model of SWB came from a single research laboratory; they typically rely on extremely small sample sizes (with per-cell samples often around 10 to 15 participants; see Yap et al. 2017); they often have effect sizes that could be considered implausibly large; and there have been

increasing numbers of failures to replicate the basic findings from the original studies. For instance, in the original small sample study showing that asking about one life domain (in this case, dating) before asking about general life satisfaction affected the size of the correlation between the two items, the correlations in the condition where general life satisfaction was asked first were much smaller than those that are typically found. These correlations ranged from −.12 to .16 (Strack et al. 1988), whereas previous research on domain satisfaction consistently shows that satisfaction with just about any domain of life tends to correlate moderately (i.e., in the range of .2 to .5) with global life satisfaction judgments, regardless of the order in which they are assessed (Schimmack and Oishi 2005).

Indeed, years after Strack et al.'s (1988) initial item-order study, Schimmack and Oishi (2005) reviewed all of the studies that they could find that used a similar methodology, and they concluded that these initial results from Strack et al. (1988) were outliers. The average correlation between global satisfaction and narrower domain satisfactions was $r = .32$ when the global measure was asked first, compared to a just slightly larger $r = .40$ when the domain satisfaction was asked first. Importantly, the difference was even smaller when the Strack et al. (1988) studies were excluded. Moreover, Schimmack and Oishi (2005) conducted five new studies, all of which found extremely small differences across conditions. Using a different design, Lucas et al. (2018) also showed that questions asked before a life satisfaction measure do not have a strong influence on responses to that global question. Thus, although making information salient before asking questions about life satisfaction may have small effects on those judgments, the striking effects presented by Strack et al. (1988) appear to be a dramatic overestimate of the typical effect of temporarily accessible information on global well-being judgments.

The concerns about replicability extend to other evidence from the judgment model. For instance, to test whether ease of retrieval affects judgments of personality characteristics, Yeager et al. (2019) used a large, nationally representative sample to test whether being asked to list six versus 12 assertive behaviors affected people's ratings of their own assertiveness. In contrast to the original small sample studies (with *N*s with approximately eight and 20 participants per cell), the effects found in this large study ($N = 1{,}338$ for a two-cell design) were not significant. Again, independent replication that used a high-quality, large sample failed to find results that were consistent with the original findings.

The most well-supported part of the judgment model concerns the mood effects on life satisfaction judgments, which have been found in multiple studies conducted by Schwarz, Strack and their colleagues. Again, however, the original studies often have characteristics associated with low rates of replicability (e.g., small sample sizes, inconsistent measures and analytic approaches, and p-values that are close to .05), and few if any successful replications have been published. Indeed, a growing number of large-scale replications have failed to find effects consistent with the original results.

For instance, Schwarz's (1987) study, which purported to show that finding a dime can affect life satisfaction ratings, had only eight participants in each of the

two conditions, and the differences between the two conditions were not significant. In another study that focused on people's reactions to soccer games, Schwarz et al. (1987) called separate groups of respondents before and after two important soccer games, one resulting in a win for the local team, and one resulting in a tie. The authors found a significant interaction between the time (before the game versus after) and the outcome (win versus tie) when predicting life satisfaction and concluded that life satisfaction judgments were influenced by the outcome of the game. However, a closer examination of the results from this study shows that there was a cross-over interaction that was driven as much by theoretically irrelevant and unexplainable differences in life satisfaction before the game as by the differences after the game, and no follow-up comparisons were conducted. Especially considering the small sample size used in the study, the robustness of these results can be questioned.

The study that received the most attention was Study 2 from Schwarz and Clore (1983). In this study, the researchers called respondents on sunny days or rainy days and found that those contacted on rainy days reported lower life satisfaction than those contacted on sunny days. The explanation was that weather affects mood, and respondents rely on their mood when making life satisfaction judgments. However, this study had six separate conditions, each of which included approximately 14 participants. Importantly, five of six conditions were not significantly different from one another (and had means that were quite similar), and just one very large discrepancy drove the entire effect. The difference in this one 14-person cell was quite large, exceeding the decline in SWB that is typically found after the onset of a serious disability or in the first year of widowhood (e.g., Lucas 2007; Lucas et al. 2003). This effect is, arguably, implausibly large given the design of the study.

In recent years, a number of researchers have attempted to replicate this and other mood studies from the judgment model tradition with little success. For instance, Lucas and Lawless (2013) used data from over one million Americans from all 50 states assessed over a five-year period, and found almost no evidence that any form of weather affected life satisfaction judgments, despite having extremely high power to detect even exceedingly small effects. Using panel data, Feddersen et al. (2016) did find significant associations between weather and life satisfaction, but they also had extremely high power, and the effect of weather found in their study was approximately one one-hundredth the size of the effect reported in the original study; i.e., a size that is arguably inconsequential and unimportant for validity.[5] Lucas and Lawless reviewed additional evidence that calls into question the robustness of the original finding of a weather effect.

Finally, because studies like those conducted by Lucas and Lawless (2013) and Feddersen et al. (2016) were conceptual replications that relied on existing data with correlational designs, Yap et al. (2017) conducted nine new experimental studies

[5] Simonsohn (2015) calculated that a one-standard-deviation increase in sunshine was associated with .01 additional points of life satisfaction on a 0-10 scale.

using procedures very similar to the classic judgment model mood studies, but with much larger samples. Although significant effects of mood induction procedures on life satisfaction were found in some of their studies (although not for weather), these effects were substantially smaller than those reported in the original study (e.g., *d*s of .09 and .11 for the two focal outcome measures). Thus, it is not clear that these effect sizes were large enough to affect the validity of traditional global well-being measures. These findings imply that concerns raised by judgment model studies about the validity of SWB measures may not be well-founded.

3.2 Testing the validity of experiential measures

The previous sections raised concerns about the robustness of existing evidence that challenges the validity of global measures of SWB. An additional issue that should be considered when comparing different measures is whether the proposed alternatives can solve the problems that global measures are thought to have. Experiential measures, such as those that can be obtained from experience sampling methodologies or the day reconstruction method, address concerns about global measures by reducing reliance on the respondents' ability to remember and aggregate. Thus, these methods address two of the most salient concerns about global measures. However, these potential strengths do not ensure that scores from experiential measures are in fact more valid than global measures of SWB. Indeed, as was noted above, it is not even clear that the concerns that have been raised about global measures translate into demonstrable decrements in validity. Thus, issues of relative validity are empirical questions that must be assessed directly.

Unfortunately, few direct comparisons of the validity of the two types of measures exist. It seems that researchers have relied on the face validity of experiential measures as the justification for their use, without subjecting these measures to more rigorous tests of validity for the purposes of assessing individual- and population-level SWB. For instance, in introducing the day reconstruction method, Kahneman et al. (2004a) simply compared absolute sample-level frequencies and diurnal patterns of affect found using this new method to those found with the existing experience sampling method. Since the publication of this initial report, few explicit tests of validity have been reported.

Admittedly, the lack of testing is attributable in part to some uncertainty about precisely how researchers should validate experiential measures. There is no gold standard measure that can be used as a criterion, and—as I have noted—widely used global measures have been criticized. Thus, the use of experiential measures has been justified primarily based on their assumed beneficial attributes. In recent years, however, my colleagues and I have attempted to address this oversight by considering additional approaches to validating these methods.

For instance, if experiential measures like those obtained from the experience sampling method and the day reconstruction method are valid, they should provide similar results when used to assess well-being in the same sample over the same time

period. Initial investigations that compared the two methods (e.g., Kahneman et al. 2004a) used different samples and only compared sample-level statistics. Studies like this, however, provide an incomplete picture of the convergence across the two methods.

Recently, we asked two samples of students to complete a day of experience sampling (in which they reported on what they were doing and how they were feeling up to nine times per day), followed by a reconstruction of the exact same day using the day reconstruction method (Lucas et al. 2021). This allowed us to compare responses at multiple levels: the sample level (e.g., were the overall sample-level estimates of time spent in specific situations and the average affect similar across methods?), the person level (e.g., did people who reported being happy with others using one method also report being happy with others using the other method?) and the moment level (e.g., when people reported being happy at a particular moment using one method, did they report being similarly happy using the other method?). Results varied depending on the level examined.

When looking at sample-level statistics, the level of agreement was quite high. For instance, when comparing estimates of time spent in various situations, the correlations across the two methods of assessment (using situations as the unit of analysis) exceeded .95. Similarly, when looking at the average positive and negative affect reported in each situation, the correlations across methods were again quite high, ranging from .76 to .85. Thus, if the goal was to estimate sample-level statistics about these experiences, both methods provided similar information.

However, when moving to the person level of analysis, the results were more varied and problematic. First, when assessing individual differences in affect experienced over the course of the day, the two methods converged reasonably well: the between-person correlations (i.e., the correlation between individual differences in average affect estimated with the ESM and average affect estimated with the DRM) for positive and negative affect were .83 and .88, respectively. Yet reports of time spent in specific situations varied across types of situations: person-level agreement was high (ranging from .74 to .83) for reports of how much time a person spent in specific locations, like home, school or work; but it was relatively low for reports of specific activities, like commuting (r = .25), eating (r = .21), or doing housework (r = .28).

Finally, when looking within individuals over time, reports of both affect and activities diverged across methods. These analyses assessed the changes in a person's affect and activities over time and examined whether the changes measured using one method corresponded to the changes measured using the other. For within-person affect, correlations ranged from r = .34 for positive affect to r = .36 for negative affect (in Study 1; results were similar for Study 2). We calculated within-person *kappa* coefficients to assess the level of agreement for dichotomous situation variables, and the average *kappa* across situations was .34. Thus, it appears that two forms of experiential measures—experience sampling and day reconstruction—did not provide the same information when used to assess a day of experience. These results raise some concerns about the validity of these measures.

Moreover, although there is no gold standard against which to compare either global or experiential measures, it is possible to compare both sets of reports to alternative measures that have different strengths and weaknesses. Lucas et al. (2021) obtained informant reports of life satisfaction from friends and family members who knew the respondent well and used these as a criterion that could be predicted from global self-reports, experience-sampling-based reports, and day-reconstruction-based reports. Although the differences in the correlations were not significant, these associations were somewhat higher for the global measures than for the experiential measures, which suggests that at the very least, the validity of the experiential measures did not exceed the validity of the global reports when informant reports were used as a criterion (also see Hudson et al. 2020, 2017; and Hudson et al. 2019a, for similar results). An additional study that looked at the associations over time between well-being and health also found that the associations were slightly stronger for the global measures than for the experiential measures (Hudson et al. 2019b). Overall, these results suggest that there is at least some cause for concern about the validity of experiential measures; and that at the very least, it should not be assumed that such measures provide scores that are more valid than those provided by traditional global measures.

3.3 Concerns about experiential measures

The goal of this paper is not simply to defend global measures and critique experiential alternatives. Instead, the goal is to raise awareness about asymmetries in the ways these classes of measures have been evaluated. Critiques of global measures often proceed by focusing on the processes that underlie the judgments that respondents are asked to make. Experiential measures are often assumed to have advantages over global measures precisely because they do not rely on the same underlying processes. However, as the evidence reviewed in the previous section shows, these fundamental differences in these measures do not always translate into improved reliability and validity for experiential alternatives to global reports. Moreover, few researchers have considered whether there are additional psychological processes underlying these experiential measures that have the potential to systematically (and negatively) affect their psychometric properties. In this final section, I discuss some potential concerns about the processes that underlie experiential measures.

Experiential measures clearly solve two challenges of global measures of SWB: they reduce reliance on memory, and they avoid the need for aggregation on the part of the respondent. However, SWB researchers have, for the most part, failed to consider whether these measures have unique problems that global measures avoid. For instance, it is possible that the increased respondent burden associated with experiential measures affects the quality of the responses that respondents provide (Eisele et al. 2020). In addition, because of logistical challenges associated with experiential measures, these measures are often more difficult to implement

than global measures in large-scale survey work. This means that selection bias may affect studies that use experiential measures more than studies that rely on global measures, as studies that use experiential measures must often rely on weaker sampling plans.

One salient difference between experiential measures and global measures is that the former measures require respondents to answer the same question multiple times so that multiple experiences can be assessed. Moods and other evaluative experiences fluctuate from moment to moment (e.g., Epstein 1979); therefore, to assess something that is representative of a person's life, multiple moments must be aggregated. Thus, respondents must report on many such occasions, either as the occasions occur (in experience sampling), or soon after they have happened (as in the day reconstruction method). However, simply asking the same question repeatedly can change the way that respondents interpret the question, which can, in turn, affect the validity (e.g., Meade and Craig 2012).

Schwarz (1999) described a model of survey response that detailed how the specific questions that researchers ask can shape the answers that respondents provide. Although researchers may assume that respondents interpret survey questions literally, subtle features of the testing environment can influence respondents' interpretations, which can, in turn, affect their responses. As just one relevant example, Strack et al. (1991) presented two similar questions to respondents: one question about happiness and one about life satisfaction. When these two questions were presented as if they were part of two separate questionnaires, respondents provided very similar responses to the two questions. However, when the questions were presented as part of the same questionnaire, respondents appeared to interpret the questions differently, and provided more discrepant responses.[6] Specifically, Strack et al. (1991) argued that in the condition in which the two questions were presented as a part of the same questionnaire, respondents assumed that the researcher would not ask the same question twice, and therefore interpreted the two questions in subtly different ways. Schwarz (1999) noted that findings like this suggest that respondents rely on *conversational norms* to interpret questions; and one important norm is that people do not ask for redundant information.

It is not difficult to see how such a conversational norm could affect experiential measures; i.e., measures that ask respondents to answer the same question over and over again across multiple experiences. It is possible that by presenting questions in this way, respondents interpret the researchers' focus as being on the *change* that occurs from moment to moment, rather than on conditions that remain stable. Thus, respondents may be more likely to report on this change, resulting in exaggerated reports of variability.

[6] Note, however, that this study also has similar characteristics to those from the judgment model, including a very small sample size and somewhat implausible effect sizes. Therefore, even this result should be interpreted with caution until it is replicated in a large-scale, pre-registered study.

At least some evidence exists to support this possibility. In a slightly different context—i.e., an investigation into personality variability—Baird and Lucas (2011) tested whether asking the same personality questions multiple times led respondents to report more variability than they otherwise would have. All participants first provided a global evaluation of their personality in an online questionnaire. Approximately one week later, they appeared in person to complete an additional set of questionnaire measures. Participants were randomly assigned to take part in one of two conditions. In the single-role condition, participants were asked to answer a set of 50 different questions about their personality in a single role (e.g., what their personality is like as a friend, as a worker or as a student). In the multiple-role condition, participants were asked a set of 10 different questions five separate times, once for each role. The hypothesis was that the variability in the responses that participants provided (as assessed by comparing each role-specific score to the general score provided a week earlier) would be greater in the multiple-role condition than in the single-role condition.

The results supported this hypothesis. Compared to the personality scores in the multiple-role condition, the personality scores in the single-role condition were more strongly correlated with and less discrepant in an absolute sense from the global scores. Thus, it appeared that compared to the reports they gave when only asked about a single role, simply asking the same question multiple times led respondents to exaggerate the variability of their personality. Moreover, participants provided responses in the multiple-role condition that were more in line with stereotypes of how people typically behave in specific situations. For instance, participants reported greater differences between the ratings of what their extraversion was like in general and what their extraversion was like as a friend in the multiple-role condition than in the single-role condition. Although it is not clear whether such effects occur in standard experiential measures of well-being, Lucas et al. (2021) reported suggestive evidence that this could be the case, especially for the day reconstruction method. Thus, the unique features of experiential measures—especially the fact that these measures rely on repeated questions—may affect the validity of the scores that result when these measures are used.

These results do not directly address the validity of experiential measures of SWB. However, like the judgment model studies of global reports, they provide insight into the processes by which participants create and report responses to the questions that are posed. Baird and Lucas's (2011) study suggests that simply asking the same question over and over may communicate to respondents that the researchers are more interested in change than in stability. This may, in turn, lead to responses that exaggerate the impact of changing situational factors on the resulting aggregate ratings. Indeed, experiential measures like those obtained from the day reconstruction method are often more strongly correlated with situational variables assessed at the same time and using the same method than they are with other major life circumstances that are often correlated with more global reports (Lucas et al. 2021). Although it is possible that this is because these situational factors truly do affect momentary experiences, it is also quite possible that these associations appear

because the use of experiential measures makes these situational factors especially salient to respondents. Again, more work that directly evaluates the validity (and the relative validity compared to global reports) of experiential measures is needed.

4 Summary

I want to be clear that in reviewing this research, I am not making the claim that the validity of global reports of well-being clearly exceeds that of experiential measures.[7] Indeed, there are many remaining questions about the validity of both types of measures, and much additional research should be done to clarify these issues. However, many prior reviews of the literature on the measurement of well-being have made two claims that I believe should be reconsidered. First, these reviews have suggested that the attention-grabbing findings from the judgment model tradition show that global measures of SWB are substantially affected by irrelevant features that reduce the reliability, stability and validity of the measures of the responses that lead to the results. The first goal of this paper was to review this evidence with a critical eye. A careful look at this literature suggests that the initial studies and results have characteristics that should raise questions about their replicability, and an increasing number of replications have failed to obtain results that were anywhere near the size of the original studies. Although the processes identified by judgment model researchers may influence well-being judgments, the bulk of the evidence suggests that the effects are small, and are unlikely to substantially affect the validity of the resulting measures.

The second claim that prior critiques of global SWB measures often made was that by solving the problems of memory and aggregation, experiential measures may represent a more valid choice for measuring SWB. Although experiential measures clearly do have these benefits, these features do not guarantee that the measures are more valid than alternatives. For one thing, the memory and aggregation problems may not have been so bad in the first place. In addition, experiential measures may have their own problems that global measures do not have. In this paper, I argued that researchers should not assume the validity of experiential measures, and that their validity must be examined empirically. I also highlighted some initial evidence that should raise some concerns about the validity and the utility of these measures, especially given their costs and burdens.

To be sure, more work on both global and experiential measures is needed. Just because the validity of global measures is not strongly affected by the processes identified by judgment model researchers does not mean that other threats to validity

[7] Though a case has been made that even if global and experiential measures are valid indicators of somewhat different constructs, global measures like life satisfaction may capture something closer to what is typically meant by SWB or may be more useful for policy purposes (Diener et al. 2009; Frijters and Krekel 2021).

do not exist. Questions about inter- and intra-personal comparability have been raised, and the answers to these questions will affect how researchers interpret responses to SWB measures. Similarly, just because challenges to the validity of experiential measures can be proposed, this does not mean that these measures do not have substantial levels of validity. It may even be the case that for some purposes, experiential measures are more valid than global measures. Indeed, studies that include both types of measures have the potential to further our understanding of the strengths and weaknesses of each. Given the wide range of applications of SWB research, an increased focus on these issues will help the field of SWB and all other fields that rely on SWB measures.

References

Baird, B. M. and R. E. Lucas 2011. "... And how about now?": Effects of item redundancy on contextualized self-reports of personality. *Journal of Personality* 79(5): 1081–1112. https://doi.org/10/fg647f

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson and M. R. Munafò 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5): 365–376. https://doi.org/10.1038/nrn3475

Cheung, I., L. Campbell, E. P. LeBel, R. A. Ackerman, B. Aykutoğlu, Š. Bahník, J. D. Bowen, C. A. Bredow, C. Bromberg, P. A. Caprariello, R. J. Carcedo, ... and J. C. Yong 2016. Registered replication report: Study 1 From Finkel, Rusbult, Kumashiro, and Hannon (2002). *Perspectives on Psychological Science* 11(5): 750–764. https://doi.org/10.1177/1745691616664694

Deaton, A. and A. A. Stone 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

Diener, E., R. A. Emmons, R. J. Larsen and S. Griffin 1985. The satisfaction with life scale. *Journal of Personality Assessment* 49: 71–75. https://doi.org/10/fqqbmr

Diener, E., R. E. Lucas and S. Oishi 2018. Advances and open questions in the science of subjective well-being. *Collabra: Psychology* 4(1):15. https://doi.org/10.1525/collabra.115

Diener, E., R. E. Lucas, U. Schimmack and J. Helliwell 2009. *Well-being for public policy*. Oxford University Press, USA. https://doi.org/10.1093/acprof:oso/9780195334074.001.0001

Diener, E., E. M. Suh, R. E. Lucas and H. L. Smith 1999. Subjective well-being: Three decades of progress. *Psychological Bulletin* 125: 276–302. https://doi.org/10.1037/0033-2909.125.2.276

Dolan, P. and M. P. White 2007. How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science* 2(1): 71–85. https://doi.org/10.1111/j.1745-6916.2007.00030.x

Eerland, A., A. M. Sherrill, J. P. Magliano, R. A. Zwaan, J. D. Arnal, P. Aucoin, … J. M. Prenoveau 2016. Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science* 11(1): 158–171. https://doi.org/10.1177/1745691615605826

Eisele, G., H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys and W. Viechtbauer 2020. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/zf4nm

Epstein, S. 1979. The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology* 37(7): 1097–1126. https://doi.org/10/c6w73c

Feddersen, J., R. Metcalfe and M. Wooden 2016. Subjective well-being: Why weather matters. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 179(1): 203–228. https://doi.org/10.1111/rssa.12118

Fredrickson, B. L. and D. Kahneman 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology* 65(1): 45–55. https://doi.org/10/dc7rrm

Frijters, P. and C. Krekel 2021. *A handbook for wellbeing policy-making: History, measurement, theory, implementation, and examples*. Oxford: Oxford University Press.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder and J. Schupp 2019. The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik* 239(2): 345–360. https://doi.org/10.1515/jbnst-2018-0022

Hudson, N. W., I. Anusic, R. E. Lucas and M. B. Donnellan 2020. Comparing the reliability and validity of global self-report measures of subjective well-being with experiential day reconstruction measures. *Assessment* 27(1): 102–116. https://doi.org/10.1177/1073191117744660

Hudson, N. W., R. E. Lucas and M. B. Donnellan 2017. Day-to-day affect is surprisingly stable: A 2-year longitudinal study of well-being. *Social Psychological and Personality Science* 8(1): 45–54. https://doi.org/10.1177/1948550616662129

Hudson, N. W., R. E. Lucas and M. B. Donnellan 2019a. A direct comparison of the temporal stability and criterion validities of experiential and retrospective global measures of subjective well-being. Manuscript in Preparation, Dallas, TX: Southern Methodist University.

Hudson, N. W., R. E. Lucas and M. B. Donnellan 2019b. Healthier and happier? A 3-year longitudinal investigation of the prospective associations and concurrent changes in health and experiential well-being. *Personality and Social Psychology Bulletin* 45(12): 1635–1650. https://doi.org/10.1177/0146167219838547

Hurka, T. 2014. Objective goods. In *The Oxford Handbook of Well-Being and Public Policy*, eds M. D. Adler and M. Fleurbaey, 379–402. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199325818.013.12

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS Medicine* 2(8): e124. https://doi.org/10/chhf6b

Ito, T. A. and J. T. Cacioppo 1999. The psychophysiology of utility appraisals. In *Well-being: The Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener and N. Schwarz, 470–488. New York, NY: Russell Sage Foundation.

John, L. K., G. Loewenstein and D. Prelec 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5): 524–532. https://doi.org/10/f33h6z

Kahneman, D. 1999. Objective happiness. In *Well-being: The foundations of hedonic psychology*, eds D. Kahneman, E. Diener and N. Schwarz, 3–25. New York, NY: Russell Sage Foundation.

Kahneman, D., B. L. Fredrickson, C. A. Schreiber and D. A. Redelmeier 1993. When more pain is preferred to less: Adding a better end. *Psychological Science* 4(6): 401–405. https://doi.org/10/djnzpd

Kahneman, D., A. B. Krueger, D. A. Schkade, N. Schwarz and A. A. Stone 2004a. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306(5702): 1776–1780. https://doi.org/10.1126/science.1103572

Kahneman, D., A. B. Krueger, D. A. Schkade, N. Schwarz and A. A. Stone 2004b. Toward national well-being accounts. *American Economic Review* 94(2): 429–434. https://doi.org/10.1257/0002828041301713

Kahneman, D. and J. Riis 2005. Living, and thinking about it: Two perspectives on life. In *The science of well-being*, eds F. A. Huppert, N. Baylis and B. Keverne, 285–304. New York: Oxford University Press, USA. https://doi.org/10.1093/acprof:oso/9780198567523.003.0011

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3): 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Krueger, A. B., D. Kahneman, D. A. Schkade, N. Schwarz and A. A. Stone 2009. National time accounting: The currency of life. In *Measuring the subjective well-being of nations: National accounts of time use and well-being*, ed A. B. Krueger, 9–86. Chicago: University of Chicago Press.

Lucas, R. E. 2007. Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies. *Journal of Personality and Social Psychology* 92(4): 717–730. https://doi.org/10.1037/0022-3514.92.4.717

Lucas, R. E. 2018. Reevaluating the strengths and weaknesses of self-report measures of subjective well-being. In *Handbook of Well-Being*, eds E. Diener, S. Oishi, and L. Tay. Salt Lake City: DEF Publishers.

Lucas, R. E., A. E. Clark, Y. Georgellis and E. Diener 2003. Reexamining adaptation and the set point model of happiness: Reactions to changes in marital status. *Journal of Personality and Social Psychology* 84(3): 527–539. https://doi.org/10/d3pbg4

Lucas, R. E., V. A. Freedman and J. C. Cornman 2018. The short-term stability of life satisfaction judgments. *Emotion* 18(7): 1024–1031. https://doi.org/10.1037/emo0000357

Lucas, R. E. and N. M. Lawless 2013. Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments. *Journal of Personality and Social Psychology* 104(5): 872–884. https://doi.org/10.1037/a0032124

Lucas, R. E., S. Oishi and E. Diener 2016. What we know about context effects in self-report surveys of well-being: Comment on Deaton and Stone. *Oxford Economic Papers* 68: 871–876. https://doi.org/10/gh5sxc

Lucas, R. E., C. Wallsworth, I. Anusic and M. B. Donnellan 2021. A direct comparison of the day reconstruction method (DRM) and the experience sampling method (ESM). *Journal of Personality and Social Psychology* 120(3): 816–835. https://doi.org/10.1037/pspp0000289

McCarthy, R. J., J. J. Skowronski, B. Verschuere, E. H. Meijer, A. Jim, K. Hoogesteyn, R. Orthey, O. A. Acar, B. Aczel, B. E. Bakos, F. Barbosa … and E. Yıldız 2018. Registered Replication Report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science* 1(3): 321–336. https://doi.org/10.1177/2515245918777487

Meade, A. W. and S. B. Craig 2012. Identifying careless responses in survey data. *Psychological Methods* 17(3): 437–455. https://doi.org/10.1037/a0028085

Mehl, M. R. and T. S. Conner, eds. 2013. *Handbook of Research Methods for Studying Daily Life*. New York, NY: Guilford Press.

National Academy of Sciences 2013. *Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*. National Academies Press. https://doi.org/10.17226/18548

O'Donnell, M., L. D. Nelson, E. Ackermann, B. Aczel, A. Akhtar, S. Aldrovandi, N. Alshaif, R. Andringa, M. Aveyard, P. Babincak, N. Balatekin … and M. Zrubka 2018. Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science* 13(2): 268–294. https://doi.org/10.1177/1745691618755704

OECD 2013. *OECD Guidelines on Measuring Subjective Well-being*. OECD Publishing, Paris. https://doi.org/10.1787/9789264191655-en

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716. https://doi.org/10.1126/science.aac4716

Redelmeier, D. A. and D. Kahneman 1996. Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66(1): 3–8. https://doi.org/10/czjgb4

Redelmeier, D., J. Katz and D. Kahneman 2003. Memories of colonoscopy: A randomized trial. *Pain* 104(1-2): 187–194. https://doi.org/10.1016/S0304-3959(03)00003-4

Robinson, M. D. and G. L. Clore 2002. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin* 128(6): 934–960. https://doi.org/10.1037/0033-2909.128.6.934

Ryan, R. M. and E. L. Deci 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55(1): 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryff, C. D. 1989. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology* 57(6): 1069–1081. https://doi.org/10.1037/0022-3514.57.6.1069

Schimmack, U. and S. Oishi 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. 1987. *Stimmung als Information: Untersuchungen zum Einfluss von Stimmungen auf die Bewertung des eigenen Lebens*. Berlin: Springer.

Schwarz, N. 1999. Self-reports: How the questions shape the answers. *American Psychologist* 54(2): 93–105. https://doi.org/10/fqrx56

Schwarz, N., H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka and A. Simons 1991. Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology* 61(2): 195–202. https://doi.org/10.1037/0022-3514.61.2.195

Schwarz, N. and G. L. Clore 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* 45(3): 513–523. https://doi.org/10.1037/0022-3514.45.3.513

Schwarz, N. and F. Strack 1999. Reports of subjective well-being: Judgmental processes and their methodological implications. In *Well-being: The foundations of hedonic psychology*, eds D. Kahneman, E. Diener, and N. Schwarz, 61–84. New York, NY: Russell Sage Foundation.

Schwarz, N., F. Strack, D. Kommer and D. Wagner 1987. Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology* 17(1): 69–79. https://doi.org/10.1002/ejsp.2420170107

Simmons, J. P., L. D. Nelson and U. Simonsohn 2011. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11): 1359–1366. https://doi.org/10/bxbw3c

Simonsohn, U. 2015. Small telescopes: Detectability and the evaluation of replication results. *Psychological Science* 26(5): 559–569. https://doi.org/10.1177/0956797614567341

Strack, F., L. L. Martin and N. Schwarz 1988. Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology* 18: 429–442. https://doi.org/10.1002/ejsp.2420180505

Strack, F., N. Schwarz and M. Wänke 1991. Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition* 9(1): 111–125. https://doi.org/10.1521/soco.1991.9.1.111

Testoni, S., L. Mansfield and P. Dolan 2018. Defining and measuring subjective well-being for sport policy. *International Journal of Sport Policy and Politics* 10(4): 815–827. https://doi.org/10.1080/19406940.2018.1518253

Tully, S. and T. Meyvis 2016. Questioning the end effect: Endings are not inherently overweighted in retrospective evaluations of experiences. *Journal of Experimental Psychology: General* 145(5): 630–642. https://doi.org/10.1037/xge0000155

Yap, S. C. Y., J. Wortman, I. Anusic, S. Glenn, L. D. Scherer, M. B. Donnellan and R. E. Lucas 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

Yeager, D. S., J. A. Krosnick, P. S. Visser, A. L. Holbrook and A. M. Tahk 2019. Moderation of classic social psychological effects by demographics in the U.S. Adult population: New opportunities for theoretical advancement. *Journal of Personality and Social Psychology* 117(6): e84–e99. https://doi.org/10.1037/pspa0000171