NOBASCHOLAR

Reevaluating the Strengths and Weaknesses of Self-Report Measures of Subjective Well-Being

By Richard E. Lucas, Michigan State University

Citation:
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. DOI:nobascholar.com

Abstract:

Subjective well-being, by definition, is a construct that relies on people's *subjective* evaluation of their lives as a whole. The subjective nature of the construct makes self-report a natural method for assessing SWB. However, the psychometric properties of these measures still must be questioned and assessed, both using traditional techniques for evaluating reliability and validity, and using more targeted approaches for understanding the processes by which respondents make these judgments. This chapter reviews the evidence about the nature and psychometric properties of self-report measures of SWB, including traditional global measures of life satisfaction and positive and negative affect, and more recently developed experiential measures including experience sampling and the day reconstruction method. Although specific questions remain about the processes underlying SWB judgments and certain limitations exist, research shows that the psychometric properties of these measures are generally quite good.

Keywords: Subjective Well-Being, Self-Report, Measurement

There are many reasons why applied and theoretical researchers seek simple, well-validated measures of the quality of a person's life as a whole. Such measures can help clarify which life circumstances or individual-level characteristics are critical ingredients of "the good life," and these findings can in turn reveal fundamental features of human nature. For instance, external circumstances that are reliably associated with differences in well-being could provide information about *basic needs* that must be satisfied for humans to flourish (Baumeister & Leary, 1995; Wilson, 1967). Similarly, individual-level characteristics or personality traits that reliably predict well-being may point towards critical skills or approaches to life that foster high levels of well-being. Such findings could be used by applied researchers at the individual level to guide personal life decisions and at the societal level to guide public policy (Diener, Lucas, Schimmack, & Helliwell, 2009).

A minimal amount of introspection reveals, however, that different life circumstances are likely to affect different people in different ways. Although strong social relationships and a meaningful career may both, on average, contribute to a good life, it also seems likely that the value that people place on these two characteristics will vary across individuals; some people may be willing to sacrifice some degree of relationship quality to obtain a successful career, whereas others may make the opposite choice. Thus, to capture one's overall quality of life, some amount of subjectivity must be incorporated into this evaluation.

Subjective well-being (SWB) is a construct that focuses explicitly on these subjective evaluations. Indeed, one simple, over-arching definition of SWB is that it represents "a person's subjective evaluation of the quality of life as a whole." This definition acknowledges that the subject's own evaluation is prioritized, which means that different people can evaluate the same objective life circumstance in different ways and can weight different life domains differently. In addition, the focus on "life as a whole" means that the totality of one's experiences and circumstances add up to an overall sense of quality of life, and that different life circumstances may balance each other out in their impact on the overall evaluation. SWB can be contrasted with *objective list* definitions of well-being, which typically identify a limited set of measurable dimensions that must be considered, without input from the subject himself or herself, to obtain

an overall evaluation of the quality of life (see Diener et al., 2009, for a discussion). SWB researchers typically object to such objective list approaches by pointing out both that it is difficult, if not impossible, to develop uncontroversial lists of important domains, and that empirical evidence suggests that people do weight domains differently.

One implication of this focus on subjective evaluations is that the study of SWB appears, necessarily, to require the use of self-report measures. If researchers wish to capture a person's own evaluation of his or her life, then it is difficult to think of alternatives to self-report measures. Unfortunately, self-reports have known limitations (Lucas & Baird, 2006; Paulhus & Vazire, 2007), which can lead to concerns about the psychometric properties of existing self-report well-being measures. Thus, it is especially important to consider the evidence that exists for the reliability and validity of subjective well-being measures. Because there is no gold-standard measures against which self-reports can be compared (as there might be for other constructs that can be assessed by self-reports, like height or one's grade point average), the process of validating measures of SWB can be challenging. In general, the process proceeds by considering evidence for reliability and construct validity, and also by addressing specific theoretical questions about how SWB evaluations are made. In this chapter, I review the evidence for the reliability and validity of SWB measures, along with the critiques of these measures and the alternatives that have been proposed.

Self-Report Approaches to Measuring Subjective Well-Being

Many different approaches to evaluating SWB have been developed, with the simplest consisting of brief, self-reported evaluations of the quality of one's life as a whole. For instance, single-item life satisfaction measures that ask "All things considered, how satisfied are you with your life as a whole" are often included in large-scale panel studies and national surveys. Multiple-item versions of these measures have also been developed, typically with similarly worded questions that tap into this overall sense that one's life is going well (e.g., Diener, Emmons, Larsen, & Griffin, 1985; Lyubomirsky & Lepper, 1999).

Diener (1984) noted that these judgments of life satisfaction form a class of measures that he referred to as being "cognitive" in nature. This term reflects the fact that answering such questions requires respondents to reflect on their lives and consciously select a response that best reflects their overall evaluation. Diener noted that such cognitive judgments could be contrasted with a separate class of evaluations that are based in respondent's typical affective experiences (e.g., Watson, Clark, & Tellegen, 1988). The reasoning behind the use of such measures is that those whose lives are going well should typically experience high and frequent levels of positive affect, along with low and less frequent experiences of negative affect. Thus, an overall evaluation of life could be obtained by capturing a person's typical levels of affect. Diener noted that these different components of SWB may ultimately result from different causes, and thus, should be assessed separately to obtain a more complete evaluation of one's life as a whole (though Busseri & Sadava, 2011, noted that there is some amount of conceptual ambiguity in this tripartite model of SWB that could be clarified with additional theoretical precision; also see Schimmack, 2008).

Traditionally, both the cognitive and affective components of SWB were assessed using global, retrospective (or "evaluative") measures that required respondents to think back on their lives and the characteristics of those lives to come up with judgments of life satisfactions or ratings of the frequency or intensity of affective experiences over some extended period (Watson et al., 1988). However, there is reason to believe that such retrospective evaluations of affective experience may not accurately capture the actual experiences that people have in their daily lives. For instance, Robinson and Clore (2002) noted that different processes are likely used to make judgments about one's current affective experience, one's very recent affective experiences, and one's affective experience from far in the past. Specifically, they argued that people can typically report with some degree of ease about their current affective experiences simply through careful introspection, but that past experiences are rated by accessing episodic and semantic knowledge. The more recent the experience, the more likely respondents are to rely on episodic memories of the actual experience; the more distant, the more likely respondents are to rely on semantic knowledge or stereotypes about how one *typically* feels in similar situations. Thus, they argued, different time lags lead to differential accuracy, if the criterion is the actual emotional experiences that would have been reported in the moment. In addition, some research suggests that when asked to aggregate over extended periods of time, certain biases affect this aggregation process, leading to global evaluations that do not match with the actual experiences as it occurred (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Redelmeier & Kahneman, 1996).

Because of the potential for inaccurate reconstructions of one's affective experience at some earlier point in time, some researchers have suggested that we should distinguish between measures that focus on retrospective evaluations and those that capture people's emotional reactions to life as it is lived (Kahneman

& Riis, 2005). For instance, it is possible to use *experience sampling methods*, where respondents are signaled multiple times per day (typically, using mobile phones) and asked to complete surveys about what they were doing, who they were with, and how they were feeling (Mehl & Conner, 2013). Kahneman (1999) argued that the benefits of these momentary---or *experiential*---reports (namely the fact that they avoid problems with inaccurate memories and aggregation biases) makes them closer to a measure of *objective happiness* than more evaluative, retrospective measures. Of course, this is an argument that must be evaluated with empirical data, as this approach may introduce new methodological problems that do not exist with more traditional retrospective methods (a concern I address below).

Finally, because methods like the experience sampling are extremely time- and resource-intensive, both for the researcher and the participant, new methodologies that combine the strengths of these experiential measures with the efficiency of more traditional questionnaire measures have been developed. For instance, Kahneman, Krueger, Schkade, Schwarz, and Stone (2004) proposed a new technique, one based in existing approaches to assessing time use, to assess affective experiences throughout the day. Their *Day Reconstruction Method* involves respondents breaking their day down into distinct episodes that represent coherent blocks of time that share some respondent-defined characteristic (e.g., "eating breakfast," "walking the dog," or "cleaning the house"). Participants are asked to list all episodes they engaged in throughout the day, to describe the features of these episodes (e.g., what the participants were doing, who they were with, where they were), and to report the affect that they experienced during each episode. The idea is that these day reconstructions, if completed relatively close in time to the actual experience, will capture relatively accurate episodic memories of the recent events. Thus, they would have many of the advantages of the experience sampling method, while only requiring a single, constrained assessment occasion. Again, as with the experience sampling method, these reasonable assumptions can be validated with empirical evidence.

The Reliability and Validity of Subjective Well-Being Measures

As I noted above, unlike certain self-report measures (such as those for height, weight, grade point average, or even for some ratings of physical health), there is no gold standard against which self-report measures of SWB can be compared. By definition, the evaluation is a subjective one, and thus, it captures internal thoughts and feelings that are not visible to outside observers. This fact, however, does not mean that self-reports themselves are unimpeachable sources of information about people's SWB. 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 (see Schwarz, 1999, for a general discussion of many of these issues). Thus, it cannot be taken for granted that any measures of SWB---self-report or otherwise---reliably and validly captures the evaluation of interest.

In general, the quality of SWB measures has been assessed using traditional approaches to understanding psychometric properties. For instance, indexes of reliability can be used to assess the extent to which measures are free from measurement error. For multiple-item measures, this is typically assessed using internal consistency coefficients, and for most measures, these coefficients tend to be high, reflecting the fact that items that are typically included in well-being measures tend to be moderately to strongly correlated (see, e.g., Diener et al., 1985). The reliability of single-item measures is slightly trickier to assess, because internal consistency coefficients cannot be used to assess reliability. The reliability of these measures is especially important, however, as single-item scales are often included in large-scale surveys. Fortunately, because the conditions of people's lives tend to be relatively stable, at least over short periods of times (e.g., most people's income, employment status, marital status, and health are stable in the short term), one would expect measures that are based, at least in part, on these circumstances, to be relatively stable, too. With this assumption in mind, short-term stability coefficients, or even longer-term stability coefficients with varying intervals can be used to assess the reliability of the measures.

For example, Lucas, Freedman, and Cornman (in press) assessed the stability over the course of a one-hour interview in a group of participants who had also been assessed numerous times over the previous six years. This research design allowed the authors to estimate stability over a very short interval, to assess the extent to which the two measures of life satisfaction were related to measures of life satisfaction assessed in prior years, and to test whether survey content administered between the two measures influenced the ratings that respondents provided. Together, these analyses provide information about the reliability of single-item measures and the extent to which any instability is due to systematic effects. Their results showed that although stability over the hour-long interview was only .62, the measure assessed at the beginning of the survey correlated just as strongly with prior years' assessments as did the measure

assessed at the end of the survey (with correlations in the range of .40-.50). Moreover, the intervening survey content appeared not to have a major impact on the final life satisfaction rating.

Other research has used multi-wave data to estimate the reliability of these single-item measures of life satisfaction (see Alwin, 2007; Kenny & Zautra, 1995 for general approaches to accomplish this goal). For instance, Lucas and Donnellan (2007) used two long-running panel studies to separate stable trait variance, reliable but slowly changing autoregressive variance, from unstable state variance in life satisfaction (which can serve as an estimate of measurement error). They estimated the reliability of life satisfaction to be approximately .67. Similarly, Lucas and Donnellan (2012) extended these analyses to four panel studies and used measures of domain satisfaction to separate reliable state variance from measurement error, and they found reliability estimates closer to .75. These estimates are also reasonably consistent with meta-analytic analyses conducted by Anusic and Schimmack (2016) and Schimmack and Oishi (2005). Together, these results suggest that single-item measures have a reasonable degree of reliability (for a more detailed review of the evidence for the reliability of single- and multiple-item measures of SWB, see Chapter 5 of Diener et al., 2009).

Assessing reliability is relatively straightforward, as this psychometric property can be described using quantitative indexes. Validity, however, is more complex and more difficult to establish in any sort of definitive way. Although validity is often simply defined as whether a test measures what it is supposed to measure, providing evidence for validity has proven to be controversial (see Borsboom, Mellenbergh, & van Heerdan, 2004, for a review and discussion). For instance, in contrast to the simple definition provided above, Messick (1995) suggested that validity could be defined as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other models of assessment" (p. 741). This definition emphasizes that validity reflects a judgment that is based on theories of the underlying construct that one hopes to assess.

Diener et al. (2009) reviewed evidence for four types of validity that could be considered when evaluating self-report SWB measures. The weakest form is *face* validity, which simply refers to whether the measure appears to measure what it is supposed to measure. Most SWB measures have high levels of face validity, though it is possible to come up with reasons why face valid measures may not be desirable or reasons why measures that lack face validity can serve as useful tools for assessing the construct. Thus, researchers tend to focus on the other three forms of validity.

Content validity refers to the extent to which the measure in question captures the breadth of the construct of interest, without including content that should theoretically be excluded. Diener et al. (2009) noted that there are two types of measures where content validity is typically a concern. First, for multifaceted measures of well-being, such as the Oxford Happiness Scale (Hills & Argyle, 2002), one could raise concerns about the inclusion of irrelevant content. For example, this scale includes the item "I think I look extremely attractive" which is not a face valid indicator of SWB. It is possible that happy people are more likely to endorse this item, regardless of their objective appearance, in which case content validity may not be affected by the item's inclusion. However, if the item truly taps the respondent's objective appearance, then the item could shift the focus of the total scale score away from its desired target.

The second type of scale for which content validity is typically a concern is for measures of the affective components of SWB. For many years, debates have taken place about the number and nature of affective dimensions that exist, and these debates shape the measures that are used to assess affective well-being. For instance, the widely used Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) does not include items assessing face valid indicators of affective well-being such as *happy* or *sad*. This is because the measure was designed to assess particular forms of positive and negative affect that are based on underlying theories about the nature of affective experiences. Researchers who use affective well-being scales should consider these existing theories and choose measures that best tap the content that they believe is theoretically most relevant (for instance, alternatives to the PANAS like the Scale of Positive and Negative Experiences [SPANE; Diener, Wirtz, Kim-Prieto, Choi, Oishi, & Biswas-Diener, 2009] may provide a better match with some measurement goals).

Convergent and *discriminant* validity are complementary forms of validity that reflect the extent to which a measure correlates strongly with other related measures (convergent validity) and correlates weakly or not at all with measures to which it should be unrelated (discriminant validity). In terms of convergent validity, it is especially important to show that different measures of the same construct cohere even when assessed using different methods of assessment. This is because correlations between measures can be inflated by shared method variance that may be independent of construct variance. Numerous studies show that standard, widely used measures of SWB tend to show reasonable levels of convergent

and discriminant validity, even when different methods of assessment are used (Lucas, Diener, & Suh, 1996). For instance, Schneider and Schimmack (2009) reported a meta-analytic average correlation of .42 between self- and informant reports of well-being.

The final form of validity, and the most complicated to assess, concerns *construct* validity (Cronbach & Meehl, 1955). This form of validity reflects the extent to which a measure behaves as it would be expected to behave, given theories about the construct itself. Thus, to test construct validity, one must have some idea about how a measure should relate to additional constructs. Diener et al. (2009) reviewed a broad range of correlates of SWB, including income, social relationships, and relevant behaviors. They conclude that widely used measures of SWB show relatively strong evidence for construct validity. For instance, there is strong reason to expect that income would be related to well-being, given that money can purchase many of the things that might be expected to cause high well-being; and indeed, income is consistently related to measures of SWB (Lucas & Schimmack, 2009). Similarly, if low levels of life satisfaction reflect a genuine evaluation of one's life as being of poor quality, one might expect scores on life satisfaction measures to predict suicidal behaviors, which appears to be the case (Koivumaa-Honkanen et al., 2001). Other studies show that domain-specific well-being measures predict relevant behaviors within those domains (e.g., job satisfaction predicts turnover; Tett & Meyer, 1993). Although this is just a brief overview of the type of evidence that can be used to assess validity, prior in depth reviews suggest that established SWB measures show an impressive degree of validity (Diener et al., 2009).

Although reliability is often a concern with single-item measures, there can also be reasons to be skeptical about the validity of such short measures. This is because psychological constructs are often complex, and it may be difficult to capture this complexity in a single item. For instance, personality traits like extraversion are often difficult to capture with a single-item because the construct itself is multifaceted. SWB constructs like life satisfaction, however, are often much simpler. It may be more reasonable to assess one's overall satisfaction with life with a single-item than it is to assess broad personality traits or attitudes. Indeed, multiple-item measures of life satisfaction often have items that are very close in meaning to one another, and those items that deviate often perform worse than the more straightforward items (Oishi, 2006). In any case, evidence suggests that when single- and multiple-item measures are compared, they tend to perform similarly. For instance, Cheung and Lucas (2014) used data from large, representative samples to compare the associations between various criterion variables and both single- and multiple-item measures, and they found that correlations tended to be very similar in size. They concluded that there was not a loss of validity when single-item measures were used.

The Judgment Model of Subjective Well-Being

One of the biggest challenges to the validity of SWB measures comes from the judgment model of SWB (Schwarz & Strack, 1999). This model starts with the reasonable assumption that when asked about their well-being, respondents do not have a response stored in memory that can simply be accessed and reported. Instead, respondents must construct a response at the time of judgment. Researchers who work from this judgment-model tradition consider the various steps that take place when such a judgment is formed, and they investigate various sources of validity and invalidity that affect each step of the process.

For instance, because respondents may not have the ability or motivation to consider all relevant domains of their lives when making a life satisfaction judgment, it is possible that they rely on various heuristics to quickly derive an appropriate response to an experimenter's question. Most famously, Schwarz and Clore (1983) suggested that respondents may rely on their current mood at the time of the judgments as a proxy for how they feel about their life overall. To provide evidence for this possibility, Schwarz and Clore (1983) manipulated mood (in one study, by having respondents write about positive or negative life events, and in a second by contacting participants on days that varied in the pleasantness of the weather) and then asked participants to report on their satisfaction with life (the experimenters made sure to present the life satisfaction question as if it were part of an unrelated task so that participants would not make a connection between the mood induction task and the well-being question). In accordance with their predictions, Schwarz and Clore (1983) found that those who were exposed to the positive mood induction procedures reported much higher life satisfaction than did those exposed to the negative mood induction procedures.

Importantly, the authors included additional conditions in their design to provide support for the idea that respondents heuristically rely on mood as a proxy for a broader life satisfaction judgment. Specifically, in additional conditions, the authors drew respondents' attention to the fact that the mood they were currently experiencing may have resulted from the prior induction. By making the cause of the mood salient, the authors argued, respondents would be less likely to rely on this mood when making the life satisfaction judgment. Again, results supported this hypothesis: the previously described mood effects on

life satisfaction judgments were eliminated when the cause of the mood was made salient. Subsequent studies replicated these results with additional mood inductions, including having one's favorite soccer team win or tie an important game or having participants complete the well-being questionnaire in a pleasant or unpleasant room (Schwarz, Strack, Kommer, & Wagner, 1987).

In a related series of studies, Schwarz, Strack, and colleagues suggested that the specific sources of information that are used when evaluating SWB can be influenced by contextual factors. Specifically, people base their judgments about the quality of their lives on different life domains, depending on which of these domains are salient at the time of the judgment. For instance, Strack, Martin, and Schwarz (1988) asked respondents about their satisfaction with life and their satisfaction with dating in one of two conditions: either with the dating question asked immediately prior to the global life satisfaction question or immediately following the question. The idea was that if satisfaction with dating was made salient, then respondents would be more likely to incorporate this specific life domain into their overall evaluation of life than if the domain had not been made accessible. Again, the results supported this prediction, with much stronger correlations between the two questions when the domain satisfaction rating was presented prior to the life satisfaction question. Schwarz and Strack (1999) review additional studies that further clarify how this information is used in global well-being judgments, along with other contextual factors (such as information about social comparison or subtle features of questionnaire design) that can impact global well-being judgments in ways that potentially affect the validity of the measures. This programmatic line of research has had an important impact on perceptions of the validity of global self-reports of SWB (e.g., Kahneman & Krueger, 2006).

Concerns About the Judgment Model

In recent years, broad concerns have been raised about standard research practices in the field, especially within (though certainly not limited to) specific domains in social psychology (Bakker, van Dijk, & Wicherts, 2012; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). These new methodological critiques suggest that certain "standard operating procedures" within the field can lead to high rates of false positive findings. These methodological practices include the reliance on very small samples, which can increase the rates of false positives that make it in to the published literature (Bakker et al., 2012); an aversion to publishing replication studies (Makel, Plucker, & Hegarty, 2012) and null results (Ferguson & Heene, 2012); and flexibility in analyses, which, when combined with natural tendencies to fall prey to confirmation bias, allow for capitalization on chance in support of a prediction (Gelman & Loken, 2013; Simmons et al., 2011). Because most of the published findings that support the judgment model were published at a time when these problematic methodological practices were the norm, it is worth revisiting the findings to determine whether the evidence is as strong as its advocates claim.

Indeed, a close look at the studies that are typically cited in support of the judgment model suggest that there is cause for concern. First, sample sizes from these studies are almost always extremely small (though for an exception, see Deaton, 2012). For instance, Yap et al. (in press) reviewed as many studies as they could find that examined the effect of mood on life satisfaction judgments and they reported that these studies had an average of just 11 participants per cell---a sample size that could only detect very large effects.

Second, the effect sizes that are typically found are arguably too large to be plausible given the hypothesized processes that underlie these effects. The effect of a mood induction procedure on life satisfaction judgments should result from the multiplicative effect of the induction on mood and the subsequent effect of mood on life satisfaction judgments. In turn, this means that the effect of a mood induction on life satisfaction judgments will necessarily be smaller than the effect of the induction on mood or mood on life satisfaction. Furthermore, one could derive plausible expectations for effect sizes from what we know about the typical effect sizes for mood induction procedures and for the association between mood and life satisfaction. As Yap et al. (in press) showed, however, effect sizes for the effect of mood induction procedures on life satisfaction judgments have typically ranged from *d*s of 1 to 2.5, which are equal to or even exceed the typical manipulation check effect sizes for mood inductions themselves (Westermann, Spies, Stahl, & Hesse, 1996)!

Finally, there is considerable evidence of analytic flexibility across studies, where different analyses are used and different effects emphasized (e.g., interaction versus simple effects) even when very similar designs were used. Together, these patterns suggest that large-sample replications would be needed before strong conclusions about the replicability and robustness of effects from the judgment-model tradition can be drawn.

Partly in response to these concerns, researchers have begun to reexamine classic effects from the judgment-model tradition using much larger sample sizes than in the originals. For example, Yap et al. (in

press) conducted nine replication studies with sample sizes that were five to ten times as large as those from the originals to examine the effect of mood on life satisfaction judgments. Effects were typically not significant, and those that were significant were considerably smaller than those from the original studies. Overall, meta-analytic estimates from the nine studies were less than one-tenth the size of the average effect from the original work (with *d*s around .15). These studies suggest that mood has only weak effects on life satisfaction judgments.

Moreover, these replication studies are consistent with other large sample studies investigating context effects on SWB judgments. For instance, Lucas and Lawless (2013) used a sample of almost one million U.S residents to examine the effect of current weather conditions on life satisfaction judgments. Even though they had considerable power to detect ever very small effects, they found no evidence that weather systematically influenced satisfaction judgments. In addition, in a study specifically designed to examine the extent to which naturally occurring fluctuations in mood were associated with fluctuations in life satisfaction judgments, Eid and Diener (2004) assessed both multiple times over a three-month period. They showed that although mood fluctuated considerably across occasions, life satisfaction judgments did not. Furthermore, the transient component of life satisfaction was only weakly associated with changes in mood, confirming with a different designs the results from Yap et al. (in press).

Additional studies have examined other context effects predicted by the judgment model and have had similar problems replicating the original findings. Most notably, Schimmack and Oishi (2005) conducted a number of studies examining item-order effects, in which the association between satisfaction with a specific domain and satisfaction with life as a whole varies depending on whether the domain was assessed prior to or after the global judgment (e.g., Strack et al., 1988). Schimmack and Oishi failed to replicate the large differences found in the original studies, and they showed in a meta-analysis that most studies that have tested for item order effects show only small differences across conditions. Schimmack and Oishi concluded that manipulating the order of items should, theoretically, not have large effects, as domains that are likely to be considered in global judgments (such as health or romantic relationships) are typically chronically salient, and thus will be included in global judgments even if the context does not make them salient. On the other hand, domains that are not chronically salient (e.g., the weather, local public transportation) are unlikely to be considered relevant for global life satisfaction, and thus will not affect satisfaction judgments even when made salient. Thus, they argued, item order effects should not be expected to be pronounced; a prediction that is supported by their large-sample studies and meta-analysis.

It is true that at least one large-sample context effect was found that had more than a modest effect size. Deaton (2012) showed that in Gallup data, asking questions about politics before a question about subjective well-being led to a moderate drop in life satisfaction, compared to a condition where these political questions were omitted. Although the mechanisms underlying this effect have not been determined (see, e.g., Lucas, Oishi, & Diener, 2016), the fact that this effect emerged in a large sample study must be acknowledged. Importantly, however, Deaton also showed that the effect was eliminated very easily simply by asking a buffer question that returned focus to the person's life before asking the well-being question. Future research can clarify when and how these effects emerge or can be prevented.

Research from the judgment model is valuable in that it pushes well-being researchers to consider the processes by which respondents formulate and report responses to questions about their SWB. The theory itself is straightforward and elegant, and it is likely that the processes identified by judgment-model researchers can impact life satisfaction judgments (as they have been shown to affect judgments in a wide range of circumstances). The important practical question, however, is not whether these processes exist, but how much they affect the psychometric properties of well-being judgments. It is possible, for instance, that the heuristics that the judgment-model researchers have identified represent just one of many processes that people can use to formulate well-being judgments, and any problems for validity that these processes introduce are outweighed by more straightforward and intuitive processes that positively impact the validity of these measures. Thus, it is important not only to look for evidence of underlying processes posited by the judgment model, but also to explicitly test the extent to which these processes undermine the validity of these measures as typically used. The research reviewed above suggests that context effects predicted by the judgment model tend to be very small; a conclusion that is supported by the considerable evidence for reliability and validity of SWB measures.

Comparing the Psychometric Properties of Global and Experiential Measures

If the two main categories of self-report measures of well-being are global measures and experiential measures, then a final issue to consider concerns the comparison between the two in terms of their psychometric properties. Remember, the idea behind the latter experiential measures is that assessing affective reactions at the time they occur (or very soon after) reduces memory problems and prevents biases that result from aggregation. Thus, it is often argued that such measures will have more desirable

psychometric properties than global measures (Kahneman, 1999). However, it is possible that these methods have unique psychometric problems that don't affect global measures. In addition, the relatively high level of respondent burden means that respondents are often assessed over a very short period of time, which can lead to concerns about the stability of results that are obtained. In other words, if respondents are asked to complete a day reconstruction method report for a single day of their lives, will the affect they experience generalize to other days and reflect a stable sense of well-being?

In their initial development efforts, the researchers who proposed the day reconstruction method did not assess long-term stability, but recent efforts have begun to provide information on this topic. For instance, Krueger and Schkade (2008) tested short-term stability and found that it was comparable to global measures of life satisfaction, even when only a single day of experiences was sampled. Similarly, both Hudson, Anusic, Lucas, and Donnellan (in press) and Hudson, Lucas, and Donnellan (2016) tested four-week stability of DRM measures, with results showing moderate stability, with correlations in the range of .40 to .60. Hudson, Lucas, and Donnellan (2016) used four waves of data, each separated by a year-long interval, to test the long-term stability of the day reconstruction methods. Again, even though just a single day of experiences was sampled, year-to-year stabilities were moderate and only slightly weaker than those from global measures like a single-item life satisfaction scale. Thus, it appears that relatively stable, trait-like measures of well-being can be obtained even from a single day's worth of reports.

These studies have also begun to compare the convergent and construct validity of global and experiential measures using large samples of participants. Anusic et al. (2017), for instance, used data from a three-wave study to examine the correlations between global and day-reconstruction-based measures of well-being and informant reports and relevant criteria (also see Hudson et al., 2016, for a similar study with similar conclusions). These studies show that global self-reports typically correlate as strongly or even more strongly with alternative measures and relevant criteria as compared to experiential measures derived from the day reconstruction method.

Given the theoretical arguments for the superiority of experiential measures over global measures, why don't experiential measures show enhanced validity when compared directly to global measures (at least in these initial studies)? One possible reason is that the two types of measures assess two distinct underlying constructs. For instance, Diener and Tay (2014) suggested that global measures are better at tapping experiences that are contextualized within a person's life schema, and that this allows these measures to predict a wider range of meaningful outcomes than more experiential measures that lack such contextualization.

Alternatively, it is possible that the unique demands of the experiential measures create specific threats to validity that do not apply to global measures. For instance, Watson and Tellegen (2002) argued that when respondents are asked to repeatedly respond to affect questions, unwanted response-style variance can, because of aggregation, get amplified relative to the variance that taps the underlying construct. Furthermore, additional research shows that asking respondents to repeatedly answer the same question over and over again can change respondent's interpretation of the questions.

For instance, Baird and Lucas (2011) examined this issue in a different domain: the repeated assessment of personality across different situations. They found that respondents reported more variability and greater impact of situational features when they were asked to report their personality multiple times in different situations than if they were asked to report on their personality in just one situation. In other words, when asked how extraverted they were in a specific situation, respondents exaggerated the effect of the situation when they were asked about multiple situations than when they were asked just about one. It is possible that experiential measures like experience sampling or day reconstruction methods communicate to respondents that they should emphasize the *changes* that occur from situation to situation rather than the *stability* that they experience. If so, this may affect associations between experiential measures and stable predictors of well-being. All in all, although experiential measures are especially useful when assessing changes in affective experiences over short periods of time, they do not have a clear advantage over global measures when used to assess stable levels of well-being. Of course, more research that directly compares the psychometric properties of these two forms of measures is needed.

Conclusion

The domain of SWB, by definition, focuses on people's subjective evaluations of their lives. Thus, there is a strong emphasis within SWB research on self-reports of the construct, as such subjective evaluations may best be assessed by asking the subject themselves. However, the subjective nature of the construct does not mean that self-reports are an unimpeachable source of information about the construct. As with all measures, the psychometric properties of these instruments must be assessed.

Fortunately, existing research suggests that SWB measures typically have desirable psychometric properties, including relatively high levels of reliability, convergent, discriminant, and construct validity. To be sure, these measures, like all self-report measures, are not perfect. Research from traditions such as the judgment model of SWB provide the means of testing some of the processes underlying well-being judgments, and when combined with more traditional approaches to evaluating psychometric properties, this research can help clarify the relative strengths and weaknesses of these approaches. Thus, research on the properties of self-report measures can not only strengthen conclusions from research that uses those methods, but can also help clarify what SWB is and how people go about evaluating their lives.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* . Hoboken, NJ: John Wiley & Sons.

Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology: Personality Processes and Individual Differences*, *110*(5), 766–781. doi:10.1037/pspp0000066

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

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. doi:10.1037/0033-2909.117.3.497

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061

Busseri, M., & Sadava, S. (2011). A review of the tripartite structure of subjective well-being: Implications for conceptualization, operationalization, analysis, and synthesis. *Personality and Social Psychology Review*, *15*(3), 290.

Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research* , *23*(10), 2809–2818.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Deaton, A. (2012). The financial crisis and the well-being of Americans 2011 OEP Hicks Lecture. *Oxford Economic Papers*, *64*(1), 1–26.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542–547.

Diener, E., & Tay, L. (2014). Review of the Day Reconstruction Method (DRM). *Social Indicators Research; Dordrecht*, *116*(1), 255–267. doi:10.1007/s11205-013-0279-x

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71–75.

Diener, E., Lucas, R. E., Schimmack, U., & Helliwell, J. (2009). *Well-being for public policy* . United States: Oxford University Press.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2009). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156.

Eid, M., & Diener, E. (2004). Global judgments of subjective well-being: Situational variability and long-term stability. *Social Indicators Research*, *65*(3), 245–277.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. doi:10.1177/1745691612459059

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences*, *33*(7), 1073–1082. doi:10.1016/S0191-8869(01)00213-6

Hudson, N. W., Anusic, I., Lucas, R. E., & Donnellan, M. B. (in press). Comparing the reliability and validity of global self-report measures of subjective well-being to experiential day reconstruction measures. *Assessment*.

Hudson, N. W., Lucas, R. E., & Donnellan, M. B. (2016). Day-to-day affect is surprisingly stable: A two-year longitudinal study of well-being. *Social Psychological and Personality Science, 8* (1), 45-54.

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

Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *The Journal of Economic Perspectives*, *20*(1), 3–24.

Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. In F. A. Huppert, N. Baylis, & B. Keverne (Eds.), *The science of well-being* (pp. 285–304). New York: Oxford University Press, USA.

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

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science*, *306*(5702), 1776–1780. doi:10.1126/science.1103572

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*(1), 52–59. doi: 10.1037/0022-006X.63.1.52

Koivumaa-Honkanen, H., Honkanen, R., Viinamaki, H., Heikkila, K., Kaprio, J., & Koskenvuo, M. (2001). Life satisfaction and suicide: A 20-Year follow-up study. *American Journal of Psychiatry*, *158*(3), 433–439. doi:10.1176/appi.ajp.158.3.433

Krueger, A. B., & Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, *92*(8-9), 1833–1845. doi: 10.1016/j.jpubeco.2007.12.015

Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod assessment* (pp. 29–42). Washington, DC: American Psychological Association.

Lucas, R. E., & Donnellan, M. B. (2007). How stable is happiness? Using the STARTS model to estimate the stability of life satisfaction. *Journal of Research in Personality* , *41*, 1091–1098.

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research*, *3*, 323–331.

Lucas, R. E., & Lawless, N. M. (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. doi:10.1037/a0032124

Lucas, R. E., & Schimmack, U. (2009). Income and well-being: How big is the gap between the rich and the poor? *Journal of Research in Personality* , *43*(1), 75–78.

Lucas, R. E., Diener, E., & Suh, E. (1996). Discriminant validity of well-being measures. *Journal of Personality and Social Psychology*, *71*(3), 616–628.

Lucas, R. E., Freedman, V. A., & Cornman, J. C. (2017). The short-term stability of life satisfaction judgments. *Emotion.* Advance online publication. doi: 10.1037/emo0000357

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. doi: 10.1093/oep/gpw023

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research; Dordrecht*, *46*(2), 137–155.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. doi:10.1177/1745691612460688

Mehl, M. R., & Conner, T. S. (Eds.). (2013). *Handbook of research methods for studying daily life* . New York, NY: Guilford Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses

and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411–423. doi:10.1016/j.jrp.2005.02.002

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

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY, US: Guilford Press.

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

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

Schimmack, U. (2008). The structure of subjective well-being. In M. Eid & R. J. Larsen (Eds.), *The science of subjective well-being* (pp. 97–123). New York: Guilford Press.

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. doi:10.1037/0022-3514.89.3.395

Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: A meta-analysis. *Social Indicators Research*, *94*(3), 363–376.

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

Schwarz, N., & Clore, G. L. (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. doi:10.1037/0022-3514.45.3.513

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). Russell Sage Foundation.

Schwarz, N., Strack, F., Kommer, D., & Wagner, D. (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. doi: 10.1002/ejsp.2420170107

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*, 429–442.

Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. *Personnel Psychology*, *46*(2), 259–293. doi:10.1111/j.1744-6570.1993.tb00874.x

Watson, D., & Tellegen, A. (2002). Aggregation, acquiescence, and the assessment of trait affectivity. *Journal of Research in Personality* , *36*(6), 589–597.

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.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, *26*(4), 557–580. doi:10.1002/(SICI)1099-0992(199607)26:4<557::AID-EJSP769>3.0.CO;2-4

Wilson, W. (1967). Correlates of avowed happiness. *Psychological Bulletin*, *67*(4), 294-306.

Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., Donnellan, M. B., & Lucas, R. E. (2016). The effect of mood on judgments of subjective well-being: Nine tests of the judgment model. *Journal of Personality and Social Psychology* . Advance online publication.