

Comparing the Reliability and Validity of Global Self-Report Measures of Subjective Well-Being With Experiential Day Reconstruction Measures

Assessment
2020, Vol. 27(1) 102–116
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191117744660
journals.sagepub.com/home/asm



Nathan W. Hudson^{1,*}, Ivana Anusic^{2,*}, Richard E. Lucas³,
and M. Brent Donnellan³

Abstract

Self-report measures of global well-being are thought to reflect the overall quality of people's lives. However, several scholars have argued that people rely on heuristics, such as current mood, when reporting their global well-being. Experiential well-being measures, such as the day reconstruction method (DRM), have been proposed as an alternative technique to obtain a potentially more accurate assessment of well-being. Across two multimethod, short-term longitudinal studies, we compared the psychometric properties of global self-reports and short-form DRM-based assessments of well-being. We evaluated their stability across one month, tested their convergent validity using self-informant agreement, and evaluated correlations with personality traits. Results indicated that global measures of well-being were more stable than DRM-based experiential measures. Self-informant agreement was also either equal across global and DRM measures or higher for global measures. Correlations with personality were similar across approaches. These findings suggest that DRM and global measures of well-being have similar psychometric properties when used to provide an overall assessment of a person's typical level of subjective well-being.

Keywords

subjective well-being, day reconstruction method, stability, reliability, validity

Subjective well-being refers to individuals' overall evaluations of the quality of their lives, as well as the balance of their day-to-day affective states (e.g., Diener, 1984). People around the world value high well-being (Diener, 2000), and they often make decisions with the goal of increasing their well-being. Moreover, governments are also becoming increasingly interested in well-being as an indicator of citizens' overall quality of life (Diener, Lucas, Schimmack, & Helliwell, 2009; Stiglitz, Sen, & Fitoussi, 2009). For example, Canada, France, and the United Kingdom have indicated plans to monitor national levels of well-being to inform public policies (Samuel, 2009; Stratton, 2010; University of Waterloo, 2011). Likewise, the U.S. Department of Health and Human Services has recently included measures of well-being in its Healthy People Initiative, which aims to improve national health (U.S. Department of Health and Human Services, 2014). In short, well-being is an important construct that is the subject of a wide range of basic and applied research.

The utility of the entire body of basic and applied research on well-being, however, hinges on a critical methodological issue: whether or not measures

of well-being have acceptable levels of reliability and validity. Specifically, some researchers have raised concerns about global measures of well-being, such as self-report life satisfaction scales, and have consequently suggested novel alternatives such as the day reconstruction method (DRM; e.g., Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004; Schwarz & Strack, 1999). However, relatively few studies have evaluated the psychometric properties of these approaches. Accordingly, the goal of the present studies was to directly compare the relative psychometric merits of global self-reports and DRM measures of well-being.

¹Southern Methodist University, Dallas, TX, USA

²Insighta Research, Toronto, Ontario, Canada

³Michigan State University, East Lansing, MI, USA

*Nathan W. Hudson and Ivana Anusic share first authorship.

Corresponding Author:

Nathan W. Hudson, Department of Psychology, Southern Methodist University, P.O. Box 750442, Dallas, TX 75275, USA.
Email: nwhudson@smu.edu

Global Measures of Well-Being

There are at least two subcomponents of well-being: global well-being and experiential well-being (Diener, Suh, Lucas, & Smith, 1999). *Global well-being* refers to individuals' top-down evaluations of the overall positivity of their lives and/or affective experiences. Global well-being is typically assessed using self-report survey-based measures with direct, face-valid items. To provide a valid response, individuals must accurately reflect on and summarize the totality of their lives and/or previous emotional experiences (A. Campbell, 1981; Schwarz & Strack, 1991).

By definition, global well-being is an *overall* evaluation of the quality of individuals' lives, and thus—barring life-altering circumstances—should remain relatively stable over time. Although the act of responding to questions about one's overall well-being might seem straightforward, the actual processes underlying such judgments has the potential to be complex (Schwarz & Strack, 1999). For example, individuals may find the task of mentally aggregating the large amount of information in their lives that is relevant to their global well-being challenging, and therefore rely on heuristics, such as contextual cues (e.g., current atypical moods; Robinson & Clore, 2002). If this is true, reports of *overall* well-being may be overly influenced by moods at the time of judgment—regardless of whether those moods accurately reflect the global quality of respondents' lives. Indeed, concerns about the role of contextual effects—such as the impact of current mood on judgments—has led some scholars to express doubts regarding the validity of global self-reports (e.g., Kahneman, 1999; Schwarz & Strack, 1999). For example, Schwarz and colleagues concluded that, “Reports about happiness and satisfaction with one's life do not necessarily reflect stable inner states” (Schwarz, Strack, Kommer, & Wagner, 1987, p. 70), and raised the possibility that there might be “little to be learned from global self-reports of well-being” because they are “too context dependent to provide reliable information about a population's well-being” (Schwarz & Strack, 1999, p. 80).

These conclusions are premised on arguments that global measures of well-being (1) exhibit low test–retest reliability and (2) can be influenced by subtle experimental manipulations (e.g., Schwarz & Strack, 1991). Global well-being should, by definition, reflect the *enduring, overall quality* of people's lives, and thus should be based on important, pervasive, and relatively stable aspects of their lives—not their immediate contextual circumstances (Diener, 1984). Thus, people's global well-being should remain relatively stable over time (A. Campbell, 1981). However, Schwarz and Strack (1991) argued that global well-being measures tend to have lower temporal reliability than would be expected of a stable construct, with maximum test–retest correlations of $r = .60$, even when assessed during the same

hour (though meta-analyses suggest that the test–retest correlation is actually higher; see Schimmack & Oishi, 2005). In addition, a few studies have found evidence that global well-being judgments appear to covary with subtle and seemingly irrelevant contextual factors, such as the weather at the time of judgment (Schwarz & Clore, 1983) or one's team recently winning a soccer game (Schwarz et al., 1987; though these findings may not be particularly replicable: Yap et al., 2017). Collectively this work suggests that global measures of well-being—which are explicitly designed to capture the *overall quality of one's life as a whole*—may be contaminated at least partially by transitory and extraneous contextual influences that should be irrelevant to judgments of the overall quality of one's life.

Alternatives to Global Measures: Experiential Measures

One solution to the concerns regarding global measures is to simply avoid them in favor of *experiential measures*, which involve repeatedly assessing momentary affective experiences. The idea is that experienced emotions—in particular those that can be described as they are happening or very soon after—are easily accessible and able to be accurately reported. Thus, assessing and aggregating experiential affect across occasions may remove participants' cognitive biases from the assessment process and provide a potentially more valid assessment of well-being (e.g., Kahneman, 1999; Kahneman et al., 2004; Robinson & Clore, 2002). Despite measuring momentary affect, one application of experiential measures is to nevertheless capture the *overall* quality of individuals' lives via aggregation (which should cause random measurement errors to cancel out and reveal people's typical, trait-like “objective” overall quality of life; Kahneman, 1999; Rushton, Brainerd, & Pressley, 1983).

Early experiential measures relied on the repeated assessment of affective experiences as they were occurring, using approaches such as experience sampling methods (ESM)/ecological momentary assessment (EMA) (Shiffman, Stone, & Hufford, 2008). However, this technique is potentially burdensome for respondents, as it requires them to carry and attend to a device that repeatedly interrupts their day to complete a survey. In addition, it is more resource intensive than simple survey studies for researchers, as researchers must find a way to ensure that respondents can be contacted to encouraged to respond to surveys in a timely fashion over the course of the study.

Because of the difficulty in implementing ESM methodologies—especially in large-scale surveys of representative populations that may be spread out geographically—Kahneman et al. (2004) developed the DRM as an alternative. In contrast to ESM, the DRM can be administered in a standard survey format in a single session. Specifically, the DRM asks participants to first divide the previous day into

specific episodes (e.g., breakfast with family, traveling to work). Participants are then instructed to recall the details of each episode and to report what they did, with whom they interacted, and how they felt. Robinson and Clore (2002) argued that people can likely report this type of momentary affect from the prior day with relatively little bias—something they appear unable to do when asked to mentally summarize and report their affect over longer periods of time (e.g., months or even weeks). Supporting this notion, Kahneman et al. (2004) provided initial evidence for the validity of the DRM by showing that recalled affective responses during key life episodes were related to other variables (e.g., age, hours of sleep) in theoretically expected ways. Moreover, subsequent research has found that, when aggregated across a day, ESM and DRM measures of affect are strongly correlated (e.g., Bylsma, Croon, Vingerhoets, & Rottenberg, 2011).

Notably, advocates of experiential measures, such as the DRM, have argued that an individual's cognitive evaluation of his or her life theoretically *should* equal the sum of his or her moment-by-moment experiences; and thus the most objective way to assess well-being (i.e., the overall quality of one's life) is by aggregating experiential ratings over a number of situations (e.g., Kahneman, 1999; Kahneman et al., 2004). According to this perspective, aggregation of DRM episode ratings should provide potentially the most valid information about individuals' overall quality of life.

There is, however, disagreement about the validity of experiential measures of well-being. For example, Tay, Chan, and Diener (2014) suggested that experiential and global measures capture different aspects of well-being, with global measures emphasizing evaluations of general quality of life, and experiential measures emphasizing actual experienced affect during specific episodes. In other words, individuals' subjective evaluations of the overall quality of their lives may provide valid information regarding how they perceive the positivity/negativity of their lived experiences, beyond what can be captured by a summation of their momentary affect. For example, working hard to serve others may foster the sense that one's life is progressing well—even if such service to others tends to generate high levels of momentary negative emotions and low levels of positive emotions. Similarly, individuals may weight their emotional experiences differentially when generating overall evaluations of their lives (e.g., individuals may report high global well-being, despite primarily feeling negative emotions [e.g., at work] because they experience primarily positive emotions “when it matters” [e.g., with family, friends]). Thus, global reports—despite not reflecting a perfect amalgam of their momentary experiences—might nevertheless provide valid information regarding the overall quality of people's lives.

Nevertheless, the suggestion that experiential measures of well-being are more valid than global ones (e.g.,

Kahneman, 1999) yields several specific, testable predictions. Namely, as aforementioned, barring life-altering events, a good measure of well-being should capture the relatively stable, overall quality of people's lives (e.g., A. Campbell, 1981; Diener, 1984). Thus, psychometrically sound measures of well-being should exhibit relatively high stability across time. Similarly, a good measure of well-being should exhibit interrater agreement and be related to theoretically relevant criterion variables, such as personality traits (e.g., Diener et al., 1999; Schimmack, Radhakrishnan, Oishi, Dzokoto, & Ahadi, 2002). Thus, claims that experiential measures of well-being are superior to global ones in terms of tapping the overall quality of people's lives imply that experiential measures should have greater temporal stability and criterion validity than global self-reports. Such claims can be directly tested by comparing the stability and validity of global and experiential measures against one another and examining the extent to which multiple observers corroborate ratings of the targets' well-being.

In sum, although there are reasons to expect that aggregated experiential assessments of well-being may potentially be more reliable, less subject to irrelevant contextual effects, and perhaps even more valid than global assessments (e.g., Kahneman, 1999), it remains an open empirical question as to whether these DRM-based measures actually have these desirable psychometric properties (Diener & Tay, 2014). In other words, it is unclear whether DRM measures are truly a superior alternative to shorter, easier-to-administer global measures (e.g., Kahneman et al., 2004). Indeed, Diener and Tay (2014) emphasized the need for greater understanding of reliability, stability, and criterion validity of the DRM. Importantly, they called for direct comparisons of the DRM to global measures of well-being—and the main purpose of this article is to address this call.

Existing Evidence Regarding the Reliability and Validity of Well-Being Measures

Considerable research has examined the reliability and validity of global well-being measures. These studies have typically examined stability over time (e.g., Anusic & Schimmack, 2016; Lucas & Donnellan, 2007, 2012), the associations between global self-reports and alternative measures such as informant-reports (Schneider & Schimmack, 2009), or associations with other important criteria, such as objective life circumstances (Lucas, 2007) and personality (Steele, 2008). With respect to stability, constructs such as well-being, which reflect stable individual differences, should exhibit relatively high test-retest reliability coefficients. In contrast, a measure that wholly reflects state-like contextual variation will approach zero stability across increasingly long periods of time. Thus,

because well-being should reflect a relatively stable individual difference (e.g., Diener, 1984), the test–retest correlations for global and experiential measures can be used to partially assess the extent to which they validly tap well-being (i.e., the overall quality of one’s life), as opposed to irrelevant contextual effects.

In a similar vein, self–other agreement in well-being ratings and the extent to which measures of well-being correlate with theoretically relevant criterion variables can be used to evaluate the relative validity of different measures. Specifically, if measures of well-being capture “real” trait-level variation in the overall quality of individuals’ lives, this variation should be observable by both the self and external informants—producing relatively high self–other agreement. In contrast, to the extent that a measure captures only fleeting and irrelevant contextual factors to which observers do not have access (e.g., the self’s memory biases and/or fleeting moods), self–other agreement should approach zero.

Similarly, well-being should theoretically be related to a number of external criteria, including personality traits (Schimmack et al., 2002; Steele, 2008). For example, extraversion and emotional stability entail stable individual differences in the propensity to experience positive and negative emotions, respectively (Goldberg, 1993). Thus, at the very least, personality traits such as extraversion and emotional stability share partial conceptual overlap with well-being and may even produce variation in well-being (e.g., high levels of extraversion may lead to greater positive affect). Indeed, research has found that personality traits are one of the strongest and most consistent correlates of well-being (Diener et al., 1999). Therefore, to the extent that measures of well-being tap “real” variation in the positivity versus negativity of individuals’ lives, they should correlate positively with extraversion and emotional stability, respectively. In contrast, if well-being measures primarily tap random contextual/state variation unrelated to dispositional tendencies to experience different kinds of emotions, we would expect little-to-no correlation with personality traits.

There has been considerable research into the reliability and validity of global measures of well-being. With respect to reliability, meta-analyses suggest that test–retest stability in global measures of life satisfaction, for example, are approximately $r = .50$ to $.60$ over a period of one to two years (Schimmack & Oishi, 2005). With regard to validity, meta-analyses have found that self and informant ratings of life satisfaction correlate approximately $r = .35$ (Schneider & Schimmack, 2009), which is comparable to agreement estimates for personality ratings (Connelly & Ones, 2010). These correlations suggest that life satisfaction ratings reflect, to a substantial degree, one’s quality of life that can be observed by another person who is unlikely to be influenced by same contextual factors that affect self-ratings

(D. T. Campbell & Fiske, 1959). Similarly, studies have also shown that global well-being relates to theoretically relevant criteria, such as stable personality traits (Schimmack et al., 2002). For example, one meta-analysis found that global measures of well-being positively correlated most strongly with extraversion, emotional stability, and conscientiousness (Steele, 2008). These correlations may suggest that global well-being judgments do not reflect only transitory contextual effects (such as atypical mood), but that they also reflect the putative influence of relatively stable dispositional factors.

In contrast to the substantial evidence for the reliability and validity of global measures of well-being, only a few studies have investigated the psychometric properties of DRM-based experiential approaches (see, Diener & Tay, 2014). For example, Krueger and Schkade (2008) found that the two-week test–retest correlations for the DRM-based methods were $r = .68$ for positive affect and $r = .60$ for negative affect. In comparison, the global life satisfaction correlation was $r = .59$ over the same period. Second, Hudson, Lucas, and Donnellan (2017b) found that the two-year test–retest correlations for DRM positive and negative affect were $r = .45$ and $.32$, respectively. As a point of comparison, in their study, the two-year test–retest correlation for life satisfaction was $r = .50$. In one final study, Dockray et al. (2010) found that the average two-hour retest correlation of DRM affective items was $r = .71$. Although this evidence is encouraging, the time periods investigated were relatively short in two of the three studies. Furthermore, none of these studies tested convergent validity by using multiple assessment methods (e.g., observer reports).

In sum, there is compelling evidence that global measures of well-being have considerable levels of reliability and validity. In comparison, much less is known about the psychometric properties of DRM-based assessments of experiential well-being. This imbalance is significant because DRM measures have the potential to provide rich information about people’s affective experiences and—*once aggregated*—might prove to be a more valid measure of well-being (i.e., overall quality of life) than are global self-reports. DRM measures are being increasingly implemented in large-scale national surveys that might inform policy, such as the German Socioeconomic Panel (Richter & Schupp, 2015) and the American Time Use Survey (2014). Thus, it is important to have a strong understanding of their psychometric properties.

Regarding aggregation, it is important to note that people’s moment-by-moment emotions are somewhat fleeting when considering any two random time points (e.g., Epstein, 1979). However, once aggregated (e.g., across a day), emotions become increasingly stable (e.g., Diener & Larsen, 1984). Proponents of DRM have typically not specified how much emotional data need to be aggregated for experiential measures to purportedly exceed the reliability and

validity of global measures. However, because of pragmatic research constraints, data are typically collected in the DRM for only one day. Although aggregating across greater intervals of time (e.g., weeks or months) might be expected to improve the validity of DRM measures (this remains, however, an empirical question), similar logic could be applied to global measures. Namely, to the extent that global measures are contaminated by random contextual influences (i.e., error), aggregating multiple measures of global self-reports of well-being will reduce random measurement errors thereby enhancing their reliability and validity coefficients because of reduced attenuation due to measurement error.

Overview of the Present Studies

In the present studies, we collected up to three measurements of global and experiential well-being across one month. Experiential well-being was operationalized via DRM reports of affective experiences from the day prior to each measurement occasion. Thus, our methodology was consistent with the typical usage of these measures—and also provided a reasonably fair test of their respective psychometric merits: How does a single day's assessment of experiential well-being compare with a single day's assessment of global well-being?

At each measurement occasion, participants completed three measures of global well-being (life satisfaction, global positive, and negative affect) as well as DRM-based measures of experiential well-being (experiential positive and negative affect). We used these data to evaluate the reliability and validity of these different approaches to measuring well-being. To the extent that either type of well-being measure is affected by contextual influences, it should be expected to demonstrate relative temporal instability. In addition to examining reliability, we evaluated evidence of validity in two ways. First, we tested convergent validity of self-rated global and DRM measures with informant-rated well-being. To the extent that well-being ratings are heavily based on irrelevant factors, such as transitory/state-level mood,¹ we would expect self-informant correlations to be very low, because informants are unaware of targets' intrapsychic contexts at the time of judgment. High correlations would suggest that both self and informant ratings reflect a visible and more "trait-like" characteristic (D. T. Campbell & Fiske, 1959). Second, we examined the extent to which well-being measures correlate with personality traits. Given that large observed correlations between self-reported personality traits and global well-being may partially reflect common method variance, we also compared self-rated well-being with informant reports of personality. All told, information from these studies will help researchers make informed judgments about assessment of well-being in research used for both basic and applied purposes.

Method

Procedure

Studies 1 and 2 had similar procedures. Both studies were three-wave longitudinal designs, with each time point separated by two weeks. At Time 1, participants completed an in-lab survey containing personality trait and well-being measures (in Study 2, personality traits were assessed via an online pretest prior to Time 1). At Times 2 and 3, well-being measures were collected online. Participants in both studies also provided names and e-mail addresses for up to six informants who knew them well enough to rate their personality traits. Informants were e-mailed a link to an online survey in which they rated personality and well-being of the target participant.

Participants

Study 1. Participants were 658 undergraduates (502 females, 145 males, 2 other gender, and 9 unreported; age $M = 19.5$ years, $SD = 2.0$ years, $range = 18-47$ years; 73% White, 12.6% Asian, 8% Black) from Michigan State University.² A total of 441 participants were rated by at least one informant, and for those participants the average number of informants was 2.0 ($SD = 1.0$).³

Study 2. Participants were 217 undergraduates (139 females, 54 males, 1 other gender, 23 unreported; age $M = 19.8$ years, $SD = 1.7$ years, $range = 18-29$ years; 80% White, 6% Asian, 5% Black) from Michigan State University. A total of 147 participants had ratings from at least one informant, and for those targets the average number of informants was 2.3 ($SD = 1.2$).

Measures

Satisfaction With Life Scale. At each wave, participants in both studies completed the five-item Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). Items similar to "In most ways my life is close to my ideal" and "I am satisfied with my life" were rated on scale from 1 (*strongly disagree*) to 7 (*strongly agree*) and averaged to form a composite.

Single-Item Life Satisfaction. Both studies included a single-item life satisfaction measure that has commonly been included in past research and national panel studies. In Study 1, the question was "How satisfied are you with your life as a whole these days?" and in Study 2 it was "All things considered, how satisfied are you with your life?" These questions were rated on a 11-point scale ranging from 0 (*completely dissatisfied*) to 10 (*completely satisfied*). For clarity, we refer to this scale as "single-item life satisfaction" (SILS), in contrast to the five-item SWLS. The

Time-1 correlation between the SWLS and SILS was $r = .65$ and $.71$ in Studies 1 and 2, respectively.

Global Affect. In each session, participants rated how often in the past two weeks they felt each of the emotions included in the survey on a scale from 0 (almost never) to 6 (almost always). In Study 1 the positive affect items were *happy*, *satisfied*, and *a sense of meaning*. Study 2 included five additional items: *friendly*, *pleasure*, *calm*, *excited*, and *competent*. Negative affect items in Study 1 were *frustrated*, *sad*, *angry*, *worried*, *tired*, and *pain*, and Study 2 included additional three items: *impatient*, *hassled*, and *criticized*.

DRM Randomly Sampled Experiential Affect. Participants provided ratings of their experiential well-being via DRM. Specifically, we asked participants to reconstruct their previous day by dividing it into specific episodes. We then randomly selected three episodes and asked more specific questions about what they did and how they felt during the episodes. Our focus in this article was on the self-rated emotions. The emotion items were identical to those assessed in the global affect measures in the two studies, and the ratings were made on scale ranging from 0 (*did not experience feeling*) to 6 (*feeling was very important part of the experience*). We computed the mean of positive and negative emotions reported for each episode, and then averaged across the episodes to obtain a single experiential positive and negative affect estimate for each reconstructed day.

Participants rated their experiential well-being in only three randomly selected episodes (as opposed to all episodes) because rating emotions in *all* episodes can take upward of 45 to 75 minutes (Kahneman et al., 2004)—which makes it unsuitable for many research contexts, including large-scale population-based surveys that have multiple research foci (e.g., Anusic, Lucas, & Donnellan, 2017). Randomly selecting three episodes provides an unbiased estimate of participants' overall affect, while simultaneously dramatically reducing the time required to complete the DRM—which increases its feasibility for use in research contexts with tight constraints on participants' time. For example, several national surveys such as the American Time Use Survey (2014), Panel Study of Income Dynamics (University of Michigan, 2014), and German Socio-Economic Panel Study (Wagner, Frick, & Schupp, 2007) have adopted the episode sampling approach for large-scale DRM assessment (see also, Anusic et al., 2017). Moreover, recent research suggests that partial DRM measures in which participants rate three randomly selected episodes are similar in terms of stability and criterion validities, as compared with full-length DRM measures (Hudson, Lucas, & Donnellan, 2017a).

Big Five Personality Traits. Study 1 included a 50-item version of the International Personality Item Pool scale (IPIP-50;

Goldberg et al., 2006), and Study 2 included the 120-item version (IPIP-120; Johnson, 2014). At Time 1 only, participants rated how accurately statements (e.g., “I worry about things,” “I have a vivid imagination”) described them as they generally are on a 5-point scale ranging from 1 (*very inaccurate*) to 5 (*very accurate*). We reverse coded appropriate items and averaged them into the Big Five personality dimensions.

Informant Measures. The informant surveys included measures of life satisfaction (both the SWLS and SILS), trait affect, and personality traits. In general, the wording was identical in the informant- and self-rated surveys, but the informant surveys referred to the participant as the target. For example, the SILS read “All things considered, how satisfied is s/he with her/his life?”⁴ When the participant indicated his or her gender, the gender-neutral pronouns were replaced with the appropriate gender-specific pronouns.

Notably, in both studies, observers rated targets' personality traits using the IPIP-50. Thus, in Study 1, the self- and informant-report personality measures were the same. In contrast, in Study 2, self- and informant-reports of personality traits were measured using different scales (the IPIP-120 and IPIP-50, respectively). Despite both scales being subsets of the larger pool of IPIP items, the 50- and 120-item scales only shared 12 items in common. Thus, in Study 2, the self- and informant-report personality trait measures differed in not only length but also item content.

Overview

We conducted three series of analyses to evaluate the psychometric qualities of our measures of global and experiential well-being. First, we examined the stability of the measures over a four-week period. To the extent that measures are contaminated by contextual factors, rather than tapping stable individual differences (as they theoretically should), we would expect to observe relatively low stability across time. Second, we examined how strongly self-rated global and experiential well-being correlated with informant ratings of well-being. To the extent that measures reflect contextual factors, rather than stable individual differences, we would expect attenuated self-observer correlations. Finally, we compared criterion-related validity coefficients with respect to the Big Five. To the extent that a measure is strongly influenced by contextual factors, we would expect to see lower correlations with personality measures than if it were influenced by stable factors.

Analytic Model

We used a variation of a trait-state model to isolate stable trait-like variance across the study from occasion-specific

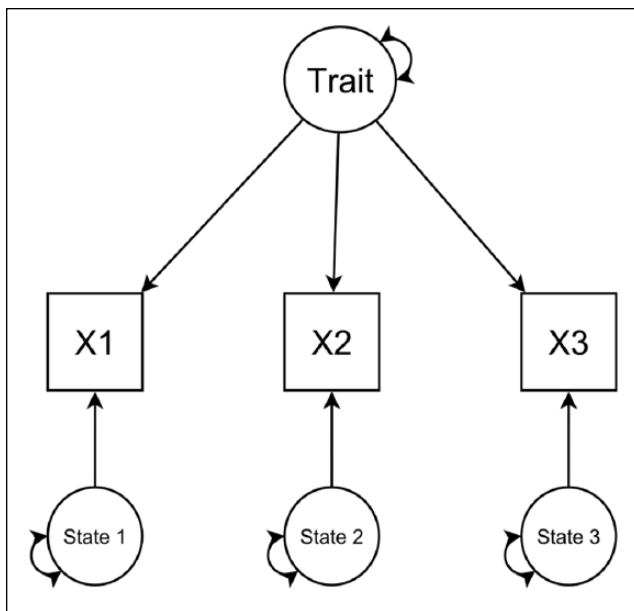


Figure 1. The trait–state model fit to three waves of data.
 Note. All state and trait loadings were constrained to 1. State variance was constrained to be equal across waves.

variance in self-rated well-being measures (Anusic, Lucas, & Donnellan, 2012; Kenny & Zautra, 2001; see Figure 1). The model assumes that there are two primary influences on a measured variable at any one point in time: trait and state factors. The stable trait factor equally influences measures at each wave, and consequently reflects all influences on the variable that are constant across the study's duration. As we did not expect that people's life circumstances would have changed greatly over the period of four weeks (on average), we assumed that the stable state factor reflected the reliable variance in well-being measures (Anusic et al., 2012; Chmielewski & Watson, 2009). In contrast, state influences change over time and are, by definition, uncorrelated across waves. In other words, these factors reflect transient contextual influences that are statistically unique to each wave, including both random measurement error and systematic factors (e.g., atypical mood) that do not carry over from one assessment to the next. For the sake of parsimony and to facilitate estimation, we constrained the amount of state variance to be equal across waves (Anusic et al., 2012; Kenny & Zautra, 2001). This constraint makes sense given that there is no theoretical reason to expect systematic changes in state variance across the intervals used in the current study. Accordingly, the proportion of observed variance in a well-being assessment allocated to either trait or state factors can be computed by dividing the amount of trait or state variance by the sum of the two estimates. We fit separate models to each of the self-rated well-being variables.⁵

Correlations With Criterion Variables. To evaluate convergent validity of self-rated well-being measures with other types of measures we extended our model to include correlations between the trait component and other criterion variables. Namely, we evaluated the extent to which the stable component of different well-being measures was related to (1) informant-rated well-being, (2) self-ratings of personality, and (3) informant ratings of personality. We fit a separate model for each of the self-rated variables correlating with each criterion variable. All analyses were performed in Mplus, with the dependency in data because of multiple informant raters providing information for some participants handled by the “cluster” function.

Results

The results of the trait-state models for each measure are shown in Table 1. Figures 2 and 3 show the absolute correlations for the trait component of self-rated well-being with informant-rated well-being and self- and informant-rated personality variables (tables of these results are available in the online supplement).⁶ The supplement also includes descriptive statistics, estimates of internal consistency, and within- and between-wave correlations of self-ratings of well-being.

Stability Over Time

Our first series of analyses allowed us to evaluate the relative impact of transient contextual effects, such as atypical mood, on different measures of well-being. As can be seen in Table 1, global measures of well-being were more stable over the course of four weeks than were the experiential DRM-based assessments in both studies. Across both studies, the trait component accounted for 80% to 81% of the variance in the SWLS, and 62% to 70% of the variance in the SILS and global positive and negative affect. The difference between the two measures of life satisfaction can be partially attributed to increased reliability of the multi-item SWLS relative to the single-item measure (Anusic & Schimmack, 2016). The experiential measures, on the other hand, showed lower stability over time: only 47%-50% of variance was stable across the four-week period.⁷ These results suggest that global well-being measures may be less influenced by transient contextual influences than are experiential DRM measures.

Convergent Correlations With Informant Ratings of Well-Being

For our next series of analyses, we examined the extent to which self- and observer-ratings of the self's well-being converged. The absolute correlations (and 95% confidence intervals) between the trait components of self-rated well-being

Table 1. Estimates of Trait and State Variances (Amounts and Percentages of Total Variance) for the Self-Rated Well-Being Variables.

	SWLS	SILS	Global PA	Global NA	Experiential PA	Experiential NA
Study 1						
Model fit						
CFI	.981	.989	.992	.976	.958	.974
RMSEA	.101	.053	.049	.086	.047	.060
Estimates						
Trait	1.07* (0.06)	2.11* (0.15)	0.66* (0.04)	0.66* (0.04)	0.72* (0.06)	0.43* (0.03)
State	0.25* (0.01)	1.31* (0.05)	0.33* (0.01)	0.33* (0.01)	0.71* (0.03)	0.48* (0.02)
Percent trait	81	62	67	67	50	47
Percent state	19	38	33	33	50	53
N	657	657	655	655	654	654
Study 2						
Model fit						
CFI	.991	.974	.950	.969	.970	.903
RMSEA	.053	.079	.114	.095	.057	.105
Estimates						
Trait	0.82* (0.09)	1.81* (0.21)	0.42* (0.05)	0.49* (0.06)	0.68* (0.10)	0.42* (0.06)
State	0.21* (0.02)	0.78* (0.06)	0.18* (0.02)	0.18* (0.02)	0.76* (0.06)	0.44* (0.04)
Percent trait	80	70	70	73	47	49
Percent state	20	30	30	27	53	51
N	217	217	217	217	216	216

Note. SWLS = Satisfaction With Life Scale; SILS = Single-Item Life Satisfaction; PA = positive affect; NA = negative affect; CFI = comparative fit index; RMSEA = root mean square error of approximation. Standard errors reported inside parentheses.

* $p < .05$.

variables and informant-rated well-being are shown in Figure 2. Informants are typically unaware of target's mood at the time they take a survey and thus do not rely on the same contextual heuristics when rating the well-being of others. Thus, to the extent that a measure is contaminated by state-level contextual factors, we would expect to see lower self-informant correlations than if both self- and informant-ratings reflected a common individual difference. As an important methodological note, it is not feasible to obtain DRM ratings from informants given likely discrepancies in how individuals will divide the day into episodes; so we used correlations between self-rated DRM with informant-rated global well-being as information about the convergent validity of DRM well-being measures.

The results showed that informant reports typically correlated more strongly with self-rated global well-being (especially life satisfaction and positive affect) as compared with experiential well-being. For example, the correlation between observer-rated global positive affect and self-rated global positive affect in Study 1 ($r = .35$, 95% CI = [.27, .43]), was larger than the correlation between observer-rated global positive affect and self-rated experiential (DRM) positive affect ($r = .25$, 95% CI = [.17, .33]).⁸ Similarly, as can be seen by comparing the extent to which the point estimates for global well-being fall outside the confidence intervals for experiential well-being in Figure 2

(and vice versa), with the exception of negative affect, global self-ratings of well-being were generally more strongly related to observer reports than were self-reported experiential measures of well-being in Study 1, but not Study 2. Thus, DRM measures certainly did not appear to exhibit *higher* convergent validity than did global measures—and if anything, DRM measures may have had lower convergent validity with informant reports than did global ones.

Correlations With Self- and Informant Ratings of Personality

For our final series of analyses, we examined the extent to which global and experiential measures were related to personality traits. We would expect to see lower correlations between self-rated personality and well-being measures that are more affected by transient contextual influences (such as atypical mood) at the time of survey compared with well-being measures that reflect overall quality of life. In addition, to rule out the potential that self-ratings of personality and well-being might be similar due to common-method variance, we replicated our analyses with informant ratings.

Figure 3 shows the absolute correlations between the trait component of self-rated well-being and self- and informant ratings of personality, along with 95% confidence intervals.⁹

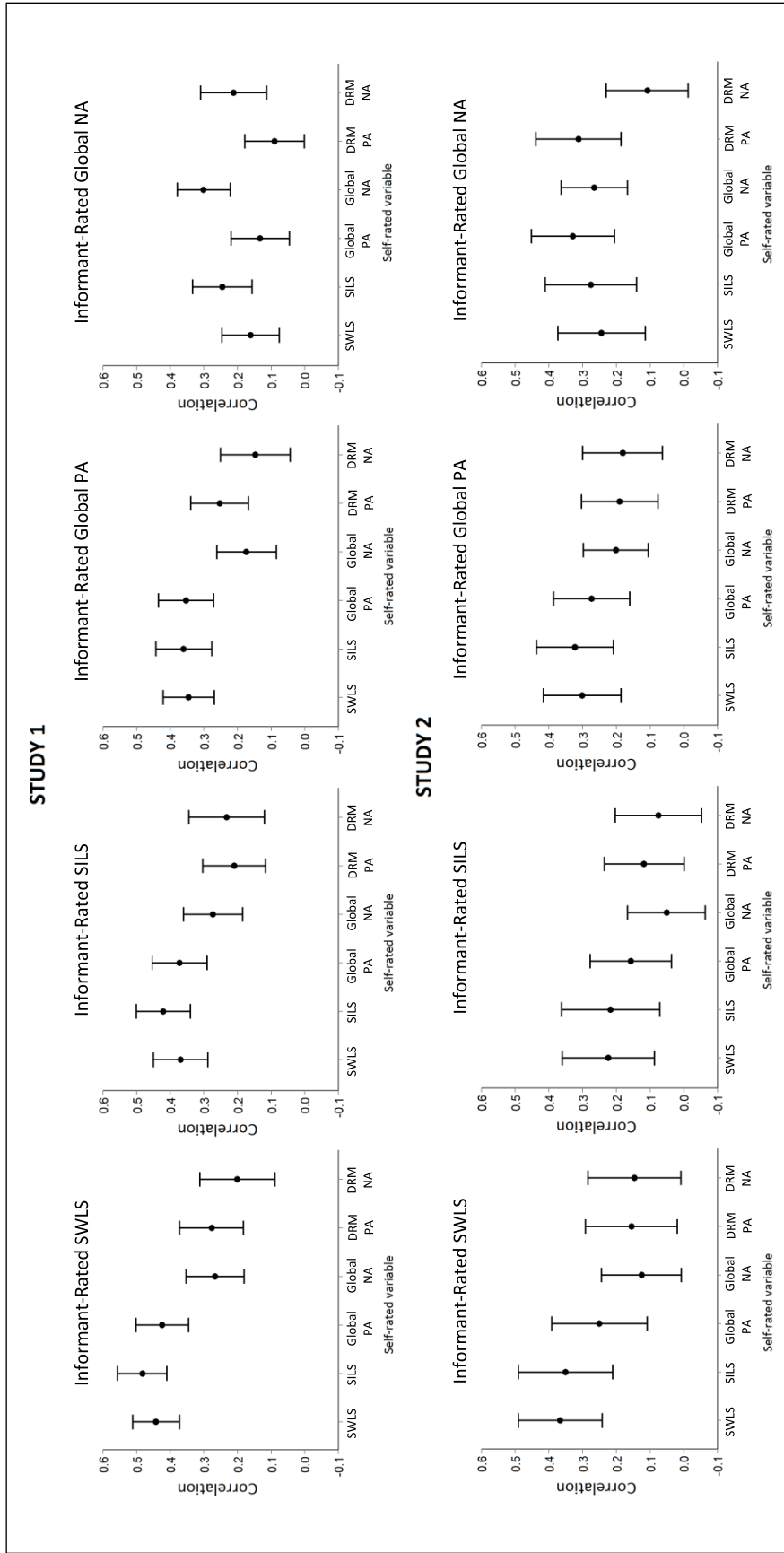


Figure 2. Absolute correlations between the self-ratings (stable trait component) and informant-ratings of well-being variables. Note. SWLS = Satisfaction With Life Scale; SILS = Single-Item Life Satisfaction; PA = positive affect; NA = negative affect. Negative correlations (e.g., with negative affect variables) have been reversed for the ease of comparison. Each plot shows correlations of all self-rated variables with a single informant-rated variable. Error bars show 95% confidence intervals around the estimated correlations.

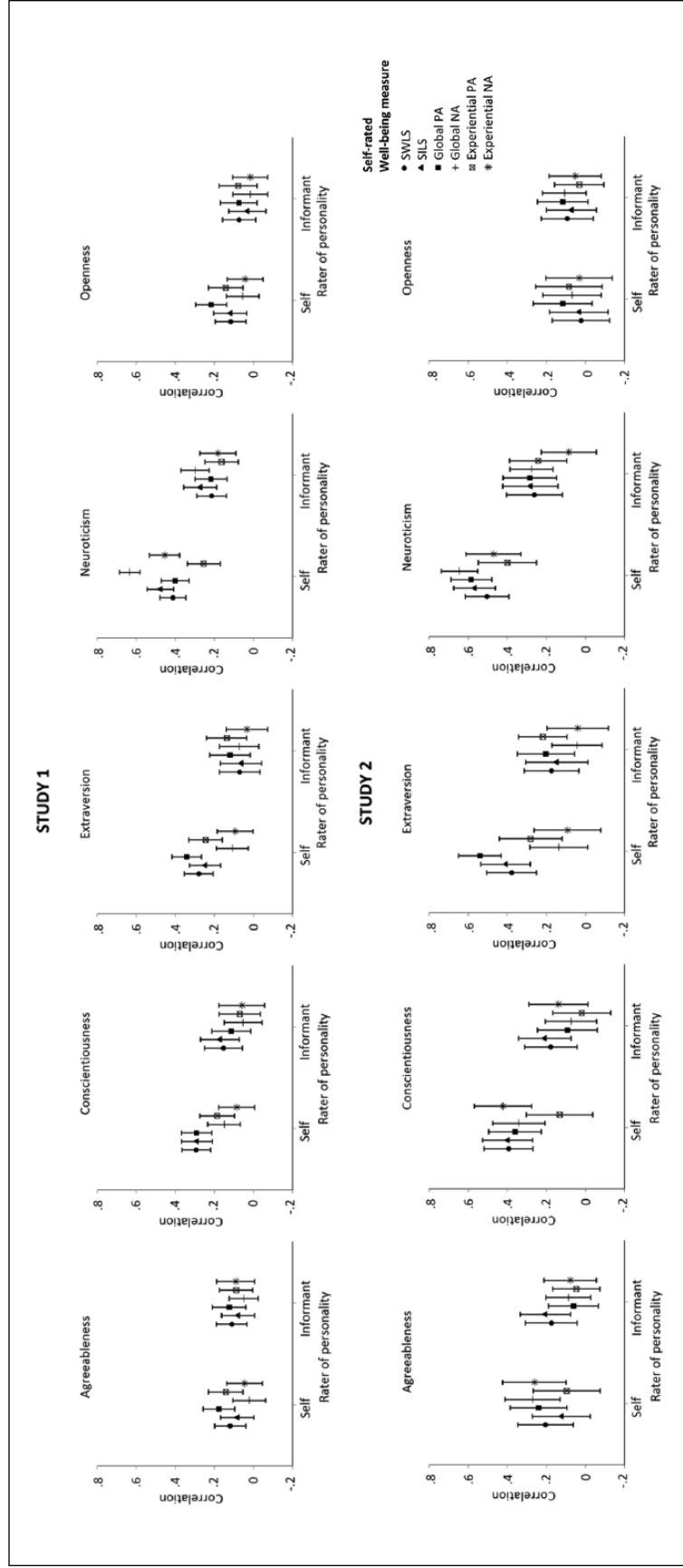


Figure 3. Absolute correlations between the stable trait component of self-rated well-being variables and self- and informant-ratings of personality. Note. SWLS = Satisfaction With Life Scale; SILS = Single-Item Life Satisfaction; PA = positive affect; NA = negative affect. Error bars show 95% confidence intervals around the estimated correlations. Global NA and experiential NA correlated negatively with all personality variables except for Neuroticism, and Neuroticism correlated negatively with all other well-being variables; we reversed these correlations for the ease of comparison.

Prior meta-analyses of well-being/personality correlations found that global measures of well-being correlate most strongly with extraversion, neuroticism, and conscientiousness (Heller, Watson, & Ilies, 2004; Steele, 2008). Our results for self-rated personality were consistent with these findings. For informant-rated personality, correlations were strongest for neuroticism. Correlations between measures of well-being and personality traits were substantially lower across informants compared with correlations within self-ratings.

As seen in Figure 3, the largest correlations between well-being and self-rated personality were observed for global measures of well-being. Correlations with DRM measures generally were lower than correlations with global measures. However, the point estimates for the global and experiential variables often fell within each other's confidence intervals—and thus were likely not substantially different from one another. Thus, the pattern was consistent with the idea that DRM measures do not have stronger, and at times may have weaker, correlations with personality (e.g., Diener & Tay, 2014). For example, the average correlations between conscientiousness, extraversion, and neuroticism and global well-being in Study 1 were $r = .28$, $.27$, and $-.45$, whereas correlations with the experiential DRM measures were $r = .19$, $.25$, and $-.24$ for positive affect and $r = -.09$, $-.09$, and $.45$ for negative affect. Results of Study 2 showed a similar pattern as Study 1, thereby strengthening our confidence in the results.

Notably, correlations between well-being and informant-rated personality were generally lower than the correlations with self-rated personality. This is consistent with observations about the influence of common method variance on the magnitude of observed correlations. However, the pattern of associations between personality and well-being remained consistent in cross-method analyses. The correlations between informant-rated personality and the DRM measures generally did not exceed correlations with global measures. In sum, DRM measures were not more strongly related to other criterion variables such as personality traits than were global measures of well-being. Indeed, the converse was true: global well-being measures generally showed equal or stronger correlations with personality measures than did DRM-based measures. This pattern persisted for correlations between self-rated well-being and informant-rated personality, indicating that it is unlikely that the observed differences between global and DRM experiential measures are due to stronger influences of common method variance on self-reports of personality and global well-being measures. Collectively, these findings provide little reason to suspect that DRM experiential measures are psychometrically superior to global measures of well-being.

Discussion

The assessment of well-being is important for both basic research and applied contexts (e.g., informing national public

policies). The most widely used method for measuring well-being is simply asking people to rate how globally satisfied they are with their lives. Such straightforward self-report measures have been shown to be reliable and valid, and they can be administered quickly. However, several scholars have suggested that the validity of global well-being measures is compromised because such measures are contaminated by irrelevant contextual factors, such as fleeting and atypical moods at the time measures are completed. To address these kinds of concerns, scholars have suggested that aggregated experiential measures, such as the DRM, may provide a more accurate assessment of well-being (e.g., Kahneman, 1999; Robinson & Clore, 2002). If true, the use of the DRM measures would result in more reliable (i.e., more stable) and valid assessments of well-being than do global self-reports.

The supposition that experiential measures have superior psychometric properties to global ones, however, has not been thoroughly tested (Diener & Tay, 2014). We therefore addressed this lacuna by comparing the reliability and validity of global well-being and DRM measures of experiential well-being in two longitudinal studies that also included informant reports. The most important finding was that DRM-based experiential measures did not have greater reliability or validity than global ones; DRM measures were, at best, equal to global ones in terms of reliability and validity—and at worst, DRM measures occasionally appeared to exhibit inferior psychometric properties, as compared with global measures. Below we comment on the broader theoretical and methodological implications of this work.

The main concern regarding global well-being measures is that people may use irrelevant contextual factors, including fleeting and atypical current moods, in addition to (or perhaps even to a greater extent than) their overall quality of life to inform their judgments (Schwarz & Strack, 1991). That is, because people do not have the time and cognitive resources to consider, identify, and average all the important aspects of their lives when forming a global judgment regarding their well-being, they tend to rely on readily accessible information at the time they are being questioned (i.e., contextual cues, mood). This possibility has important implications for researchers who wish to develop and extend theories of well-being, and for public policy makers who want to factor quality of life into decision making. In essence, some have suggested avoiding the use of global self-reports if the objective is to have a valid indication of individuals' general affective tendencies and overall quality of life. Aggregated experiential measures, such as the DRM, are purported to be more accurate because the responding task is more directed and structured for participants. Instead of making global reflections, participants merely have to recall how they felt during specific episodes from the previous day—something they appear to be able to do relatively accurately (e.g., Robinson & Clore, 2002). These ratings

can then be aggregated to provide an overall assessment of well-being.

The prediction that global self-reports of well-being are more strongly influenced by contextual factors than are aggregated experiential ones suggests that global ratings should be less stable in the short term than DRM-based measures because contextual factors tend to fluctuate from occasion to occasion. However, our findings seem to directly contradict this prediction. In our study, a multiple-item measure of life satisfaction was the most stable, followed by single-item life satisfaction and global positive and negative affect. The DRM-based experiential measures were the least stable. Thus, at least in these two studies, asking people to rate their overall well-being generated a more stable estimate of well-being than averaging affect across DRM episode ratings. This seems to suggest that global judgments reflect, to a reasonable degree, a relatively stable individual-difference construct (Schimmack & Oishi, 2005).

Beyond issues of stability, we also considered the convergence between self- and informant-ratings of well-being. Targets and informants are unlikely to be influenced by same transient contextual factors (e.g., situational cues, atypical mood). Thus, if self-ratings tend to be affected by contextual factors, such as mood at the time of judgment, correlations with informant ratings of well-being should be relatively low. Likewise, if DRM-based measures provide a more valid assessment of well-being than do self-reports, self-informant agreement should be higher for DRM-based measures than for global measures. In contrast to this prediction, higher agreement with informant-rated well-being was generally found for self-rated global well-being. This finding suggests that global well-being judgments reflect a relatively enduring individual difference rather than merely a snapshot of current contextual factors.

A final way we tested the criterion-related validity of global and experiential well-being was to evaluate their correlations with personality traits. If DRM measures captured stable aspects of one's quality of life better than did global measures, we would expect to see higher correlations between DRM-based measures and personality traits. This was not the case, as DRM correlations with personality were not consistently superior to global self-reports in this regard. As personality ratings are less likely to be affected by mood (Eid & Diener, 2004), these findings provide additional evidence that global measures are not simply proxies for contextual factors, such as current and atypical mood.

In sum, we did not find consistent evidence pointing to the superiority of DRM-based experiential assessments over global self-reports of well-being. Although we believe that our results should prove reassuring to researchers who rely on global self-reports to assess well-being, there are a number of caveats. One possible reason for lower stability and lower criterion correlations of DRM measures relative

to global measures may be that a sample of three episodes drawn from a single day may simply not be enough to obtain a reliable indicator of a person's actual quality of life. DRM measures contain two sources of unreliability when it comes to assessing well-being—unreliability at the item level (e.g., measurement error) and variation across rated episodes. It is possible that rating episodes from every day over a much longer time period (e.g., a month or even a year) would provide more accurate insight into global quality of life—although it is still unclear whether this estimate would be better than what can be obtained via global well-being measures. Unfortunately, this approach would probably be impractical for assessment in most research contexts, including national surveys, which are one of the important applications of the DRM (e.g., Kahneman et al., 2004). Moreover, the same logic regarding aggregation would apply equally to global ratings of well-being: Obtaining and averaging multiple measures of global well-being should cause random, contextual influences to mutually cancel, leading to a more accurate assessment of well-being, in the same way that obtaining more DRM ratings may lead to more accurate estimates of well-being. Thus, a “fair” comparison of the relative psychometric properties of DRM/experiential and global measures requires an identical number of measurements aggregated for both measures (e.g., both DRM and global measures assessed every day for a month and aggregated).

A second concern is that we used college students who were primarily female for our studies. Future research should evaluate whether our findings generalize to nonstudent and/or more male populations. For example, stability of global well-being judgments increases with age (Lucas & Donnellan, 2007). It is possible that DRM may also be more stable in populations of older adults if, for example, their lives are more structured on daily basis, as compared with lives of college students. Comparing stability and reliability of the DRM-based measures in different populations would provide important information for assessment of well-being at the national level (e.g., Hudson et al., 2017b). Relatedly, we used a relatively limited set of criterion related variables (i.e., personality traits). Although personality traits are one of the strongest and most consistent predictors of well-being (Diener et al., 1999), future research could explore whether our results also generalize to other criterion variables (e.g., age, income, life circumstances).

A third limitation of our study pertains to the self- and informant-ratings of personality in Study 2. We found that personality exhibited similar correlations with well-being, irrespective of whether the self- or informant ratings were used. However, we also found that, in some instances, self-rated traits had somewhat higher correlations with self-reported well-being than did the informant ratings. Although we believe that this represents common-method variance in self-report ratings, ultimately the self- and informant-report

scales in Study 2 differed both in terms of scale length (120 vs. 50 items, respectively) as well as in terms of the specific items that were included in each scale. Thus, we cannot soundly rule out the possibility that the slight differences in correlations observed in Study 2 between self- and informant-ratings are attributable to these differences in the scales, rather than the source of the rating (i.e., self- vs. informant reports; although notably, similar differences were observed in Study 1 in which the self and observer scales contained the same items). On a similar note, the observer ratings came from heterogeneous sources (e.g., parents, friends, romantic partners). This heterogeneity may have reduced the magnitude of associations between self- and observer-reports. Future research could explore whether the source of observer ratings moderates the association between self- and observer reports.

A fourth concern is that we used the overall mean of the DRM affect ratings across episodes, rather than the duration-weighted mean suggested by Kahneman et al. (2004). Weighting episodes by their duration would not be advisable here because we obtained only a random set of three episodes per person per measurement occasion; so weighting may not adequately reflect the total positive and negative emotions experienced in a day. In addition, the practice of weighting by duration has been questioned by other researchers (e.g., Diener & Tay, 2014). Nevertheless, it is possible that other ways of summarize DRM ratings may be more appropriate than the simple mean. Future research should evaluate the best practices for estimating overall well-being from momentary ratings of affect and samples of affect.

A fifth concern is that our study assessed experiential affect via DRM as opposed to ESM. The DRM entails some level of retrospective reporting, and thus may introduce memory biases that influence well-being ratings (cf. Robinson & Clore, 2002). Thus, although our primary goal was to examine specifically the psychometric properties of the DRM, our findings may not generalize to other measures of experiential affect, such as ESM (notably, however, studies suggest that daily aggregates of DRM and ESM affect correlate strongly with one another; Bylsma et al., 2011).

A sixth limitation of the present study is that we did not measure life events that might be theoretically expected to influence well-being. Thus, we were unable to compare whether global and experiential measures of well-being systematically vary with relevant life events in sensible ways. Future research might consider including measures of life events and examining the extent to which global and experiential well-being are responsive to changes in life circumstances.

Finally, although our results suggest that DRM-based assessments are not superior to global self-reports, it is important to emphasize that the DRM can still provide valuable information for researchers and policy makers alike. For example, this method can provide insight into people's time

use and satisfaction with daily activities. It can also be used to develop and test focused hypotheses about dynamics of affective experiences across situations. For example, Oishi, Kurtz, Miao, Park, & Whitchurch (2011) found that retired individuals reported higher positive affect in familiar rather than unfamiliar places, whereas working individuals reported more positive affect in unfamiliar places. These authors also found that familiarity with interaction partners correlated with ratings of positive affect for Korean, but not American participants. Similarly, Hudson, Lucas, and Donnellan (2017c) found that people's experiential affect varied as a function of the individuals with whom they were currently interacting across various situations. This type of situational sensitivity cannot be achieved with global measures.

In closing, the accurate assessment of well-being is a critical issue for psychological science—both in terms of basic research seeking to understand the processes related to and correlates of well-being, and also in terms of applied contexts, such as informing public policies. High-profile criticisms of global self-reports have motivated researchers to develop and emphasize experiential measures as alternatives to global ones. However, these approaches have not been subjected to rigorous comparative evaluations to determine how they compare with the simpler global self-report method. The current study suggests that the DRM is not superior to global self-reports. Indeed, some of the criticisms of global self-reports may have been too strident, as the empirical evidence in this report and others (e.g., Hudson et al., 2017a, 2017b) suggests that global self-report measures are a reasonably valid approach for assessing subjective well-being—at least compared with the DRM-based measures commonly used in large-scale panel studies. Tentatively, it seems that simply asking people to reflect on their lives is, in fact, an efficient and effective way to capture their overall well-being, after all.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the National Institutes of Health National Institute on Aging (No. AG040715).

Supplemental Material

Supplementary material is available for this article online.

Notes

1. Notably, in this context, “mood” refers to random, state-like contextually driven variation in people's affective experiences. In contrast, characteristic or stable “moods” (e.g., a

- person who frequently experiences negative moods) represent trait-like positive and/or negative affect.
2. Information about compliance and samples sizes at each wave is included in the online supplement.
 3. The majority of informant ratings were completed by parents and close friends (55% in Study 1, 56% in Study 2). An additional 24% of informant ratings in Study 1 and 28% in Study 2 were completed by siblings, romantic partners, and roommates. Other family members, acquaintances, and coworkers were responsible for the remaining informant reports.
 4. One exception to this wording format was that for the SWLS items informants rated the degree to which they thought the targets agreed with statements such as “In most ways my life is close to my ideal” rather than how much the informants themselves agreed with the statement “In most ways her/his life is close to her/his ideal.”
 5. Overall, the models fit reasonably well according to the criteria suggested by Hu and Bentler (1999; see Table 1). The fit of the models could generally be improved by allowing state variance to vary across waves. However, there was no systematic variability in the state component across different measures, and the interpretation of results becomes more complicated once this constraint is relaxed. For these reasons we decided to proceed with fully constrained models.
 6. The correlations for self-rated well-being from each of the three individual waves with criterion variables are reported in the online supplement. These additional correlations were generally smaller in magnitude than correlations with the trait component of self-ratings (because of the fact that each occasion included measurement error). The supplemental results can provide researchers with a sense of what should be expected in studies with only a single assessment but the pattern mirrored the trait component correlations reported here.
 7. The online supplement shows the full correlations matrix for all measures across all waves.
 8. In a model directly comparing these correlations, constraining them to be equal significantly worsened the model fit, as compared with allowing them to freely vary from one another, $\chi^2(1) = 4.09, p = .04$.
 9. Correlations between self- and informant-ratings of personality can be found in the online supplement.

References

- American Time Use Survey. (2014, July). *Well-Being Module Questionnaire*. Retrieved from <http://www.bls.gov/tus/wbm-questionnaire.pdf>
- Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *Journal of Personality, 80*, 33-58.
- Anusic, I., Lucas, R. E., & Donnellan, M. B. (2017). The validity of the day reconstruction method in the German Socio-Economic Panel Study. *Social Indicators Research, 130*, 213-232.
- 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, 110*, 766-781.
- Bylsma, L. M., Croon, M. A., Vingerhoets, A. J. J. M., & Rottenberg, J. (2011). When and for whom does crying improve mood? A daily diary study of 1004 crying episodes. *Journal of Research in Personality, 4*, 385-392.
- Campbell, A. (1981). *The sense of well-being in America*. New York, NY: McGraw-Hill.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology, 97*, 186-202.
- Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092-1122.
- Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542-575.
- Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55*, 34-43.
- 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., & Larsen, R. J. (1984). Temporal stability and cross-situational consistency of affective, behavioral, and cognitive responses. *Journal of Personality and Social Psychology, 47*, 580-592.
- Diener, E., Lucas, R., Schimmack, U., & Helliwell, J. (2009). *Well-being for public policy*. New York, NY: Oxford University Press.
- Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. E. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*, 276-302.
- Diener, E., & Tay, L. (2014). Review of the day reconstruction method (DRM). *Social Indicators Research, 116*, 255-267.
- Dockray, S., Grant, N., Stone, A. A., Kahneman, D., Wardle, J., & Steptoe, A. (2010). A comparison of affect ratings obtained with ecological momentary assessment and the day reconstruction method. *Social Indicators Research, 99*, 269-283.
- Eid, M., & Diener, E. (2004). Global judgments of subjective well-being: Situational variability and long-term stability. *Social Indicators Research, 65*, 245-277.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.
- Heller, D., Watson, D., & Ilies, R. (2004). The role of person versus situation in life satisfaction: A critical examination. *Psychological Bulletin, 130*, 574-600.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.
- Hudson, N. W., Lucas, R. E., & Donnellan, M. B. (2017a). *Comparing the temporal stability and criterion-related*

- validity of experiential and self-reported global measures of well-being. Manuscript submitted for publication.
- Hudson, N. W., Lucas, R. E., & Donnellan, M. B. (2017b). Day-to-day affect is surprisingly stable: A 2-year longitudinal study of well-being. *Social Psychological and Personality Science*, *8*, 45-54.
- Hudson, N. W., Lucas, R. E., & Donnellan, M. B. (2017c). *Are we happier with others? An investigation of the links between spending time with others and subjective well-being*. Manuscript submitted for publication.
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78-89.
- 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., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *3*, 1776-1780.
- Kenny, D. A., & Zautra, A. (2001). The trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 243-263). Washington, DC: American Psychological Association.
- Krueger, A. B., & Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, *92*, 1833-1845.
- Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science*, *16*, 75-79.
- 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*, *105*, 323-331.
- Oishi, S., Kurtz, J. L., Miao, F. F., Park, J., & Whitchurch, E. (2011). The role of familiarity in daily well-being: Developmental and cultural variation. *Developmental Psychology*, *47*, 1750-1756.
- Richter, D., & Schupp, J. (2015). The SOEP Innovation Sample (SOEP IS). *Schmoller Jahrbuch*, *135*, 389-400.
- Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*, 934-960.
- Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*, 18-38.
- Samuel, H. (2009, September 14). Nicolas Sarkozy wants to measure economic success in "happiness." *The Telegraph*. Retrieved from <http://www.telegraph.co.uk/news/worldnews/europe/france/6189530/Nicolas-Sarkozy-wants-to-measure-economic-success-in-happiness.html>
- Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, *89*, 395-406.
- Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: Integrating process models of life satisfaction. *Journal of Personality and Social Psychology*, *82*, 582-593.
- Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: A meta-analysis. *Social Indicators Research*, *94*, 363-376.
- 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*, 513-523.
- Schwarz, N., & Strack, F. (1991). Evaluating one's life: A judgment model of subjective well-being. In F. Strack, M. Argyle & N. Schwarz (Eds.), *Subjective well-being: An interdisciplinary perspective* (pp. 27-47). Oxford, England: Pergamon.
- Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In D. Kahneman, E. Diener & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61-84). New York, NY: 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 life-domains. *European Journal of Social Psychology*, *17*, 69-79.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1-32.
- Steele, P. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*, 138-161.
- Stiglitz, J. E., Sen, A., & Fitoussi, J. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Commission on the Measurement of Economic Performance and Social Progress. Retrieved from <http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report>
- Stratton, A. (2010, November 14). Happiness index to gauge Britain's national mood. *The Guardian*. Retrieved from <http://www.theguardian.com/lifeandstyle/2010/nov/14/happiness-index-britain-national-mood>
- Tay, L., Chan, D., & Diener, E. (2014). The metrics of societal happiness. *Social Indicators Research*, *117*, 577-600.
- University of Michigan. (2014). *Panel Study of Income Dynamics*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from <http://psidonline.isr.umich.edu/Guide/documents.aspx>
- University of Waterloo. (2011, October 20). *CIW releases Canada's first-ever index to measure national wellbeing*. Retrieved from <http://uwaterloo.ca/applied-health-sciences/news/ciw-releases-wellbeing-index>
- U.S. Department of Health and Human Services. (2014). *Healthy People 2020*. Retrieved from <http://www.healthypeople.gov/2020/default.aspx>
- Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch*, *127*, 139-169.
- Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., Donnellan, M. B., & Lucas, R. E. (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. doi:10.1037/pspp0000115